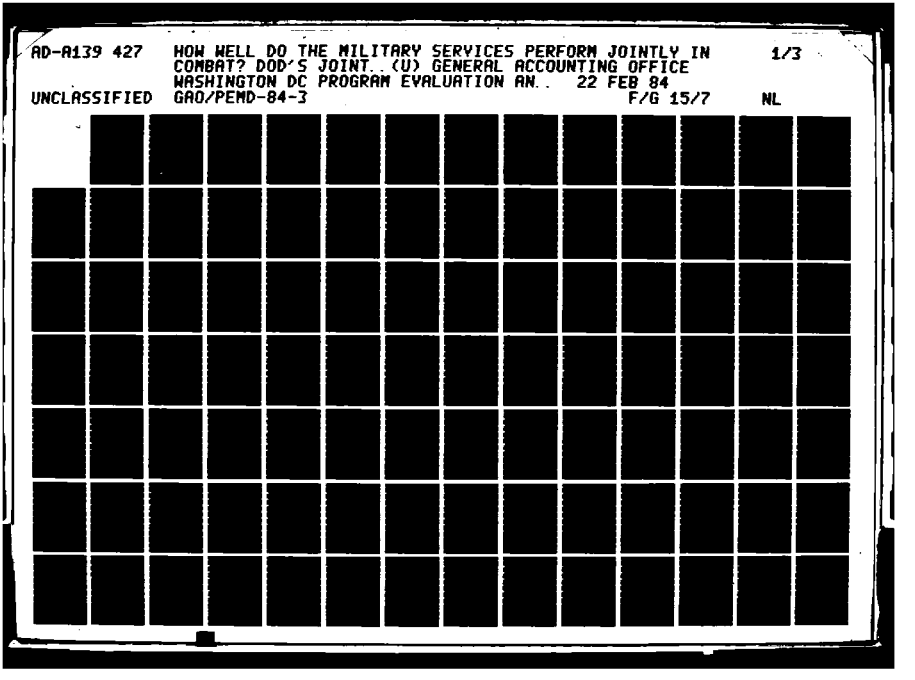
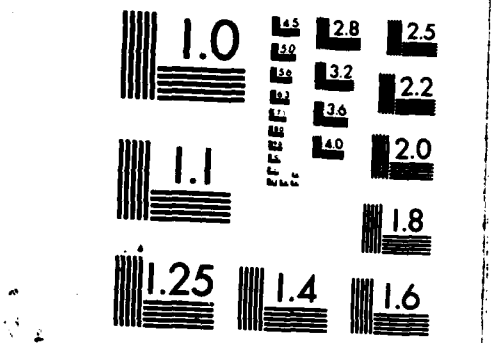

AD-A139 427
HOW WELL DO THE MILITARY SERVICES PERFORM JOINTLY IN COMBAT? DOD'S JOINT.. (U) GENERAL ACCOUNTING OFFICE WASHINGTON DC PROGRAM EVALUATION AN.. 22 FEB 84
1/3
UNCLASSIFIED
GAO/PEMD-84-3
F/G 15/7
NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

BY THE COMPTROLLER GENERAL

Report To The Honorable David Pryor United States Senate

OF THE UNITED STATES

AD A139427

How Well Do The Military Services Perform Jointly In Combat? DOD's Joint Test-And-Evaluation Program Provides Few Credible Answers

In most combat situations, DOD combines military missions and their operations in ways that transcend the boundaries and responsibilities of the individual armed services. To ascertain whether DOD's joint test-and-evaluation program produces credible information about these situations, GAO reviewed 13 JT&E's that were completed between 1972 and 1981, analyzed 3 in depth for their systemic strengths and weaknesses, and examined the history, organization, and management of the JT&E program.

In examining the test process, GAO found that the most important threats to the quality and usefulness of JT&E results are test formulations that fail to consider critical issues, test designs that set up unrealistic test conditions, test implementation that deviates from the test design, test analysis that fails to employ appropriate techniques or to control for validity, and test reports that are untimely, based on faulty interpretations, or not appropriately balanced or qualified. Some of these threats may stem from the organizational features of the JT&E program that make it dependent both on DOD's weapons developer and on the services for their expertise, planning, resources, and funding.

All classified information in this report has been deleted. The complete classified version of this report is published under the number GAO/C-PEMD-84-1.

DTIC ELECTE MAR 2 7 1984 A

DTIC FILE COPY

This document has been approved for public release and sale; its distribution is unlimited.

GAO/PEMD-84-3
FEBRUARY 22, 1984

84 03 26 05

COMPTROLLER GENERAL OF THE UNITED STATES
WASHINGTON D.C. 20548

B-207947

The Honorable David Pryor
United States Senate

Dear Senator Pryor:

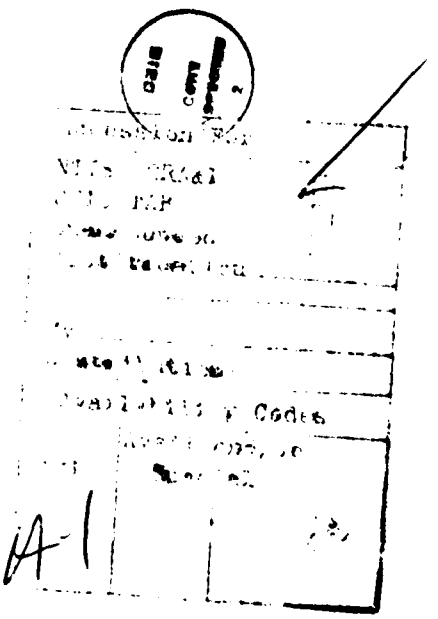
This report responds to your May 17, 1982, letter asking us to assess the quality and usefulness of the joint testing and evaluation program of the U.S. Department of Defense. This is an unclassified version of our classified report numbered GAO/C-PEMD-84-1.

Officials of the Department of Defense were asked to comment on a draft of this report. DOD elected not to comment.

As you requested, we plan no further distribution until 30 days from the date of the report unless you publicly announce this report earlier. At that time we will send copies to interested congressional committees; the Secretaries of Defense, Army, Air Force, and Navy; the Chairman of the Joint Chiefs of Staff; and the Director of the Office of Management and Budget.

Sincerely,

Charles A. Bowsher
Comptroller General
of the United States



RE: Classified References, Distribution
Unlimited
No change per Ms. Butler, GAO/Distribution
Section

COMPTROLLER GENERAL'S
REPORT TO THE HONORABLE DAVID PRYOR
UNITED STATES SENATE

HOW WELL DO THE MILITARY SERVICES PERFORM JOINTLY IN COMBAT? DOD'S JOINT TEST-AND-EVALUATION PROGRAM PROVIDES FEW CREDIBLE ANSWERS

D I G E S T

In 1971, the U.S. Department of Defense (DOD) established a joint test-and-evaluation program (JT&E). The primary purpose of the JT&E program is to find out how well the military services can perform their missions and roles in joint operations under combat conditions. JT&E is separate and apart from the testing programs of the individual armed services. The Air Force, Army, and Navy programs focus primarily on their single service operations, whereas JT&E focuses on joint operations involving more than one of the armed services.

Senator David Pryor asked GAO to review the JT&E program to ascertain whether it has produced credible information about how well military operations involving more than one of the armed services can be performed. Senator Pryor asked GAO several specific questions about the operation and management of the JT&E program and the quality and usefulness of the tests. To answer these questions, GAO (1) examined the history, organization, and management of the JT&E program, (2) reviewed the 13 JT&E's that were completed between 1972 and 1981, and (3) analyzed 3 of these JT&E's in depth for their systemic strengths and weaknesses. The judgmental selection of the 3 tests was based on the following criteria. First, only recently completed tests were considered. Second, tests were sought that would illustrate several types of JT&E, differing in purpose, requestors, and the participation of the armed services. This digest summarizes GAO's observations on the overall management of the JT&E program and on the quality and usefulness of the 3 tests that were analyzed. The final section summarizes GAO's responses to the questions posed by Senator Pryor.

JT&E PROGRAM ORGANIZATION AND MANAGEMENT

The JT&E program was established in 1971 following the recommendations of a blue-ribbon

panel that such an activity be established at a higher-than-service level in DOD. In accordance with DOD Directive 5000.3, the office of the Director for Defense Test and Evaluation (DDT&E) is responsible for the JT&E program. This office reports to the Under Secretary for Defense Research and Engineering, the weapons-developer organization for the Office of the Secretary of Defense (OSD). Few staff members are assigned to JT&E functions: approximately seven persons in the office of the DDT&E spend less than 30 percent of their time on JT&E. They are, with one exception, career military officers on rotation for two or three years from their services. They are not chosen primarily for their testing expertise. (pp. 7-8)

Although the Defense Test and Evaluation office has primary responsibility for the JT&E program, it has limited resources (staff, funding, test equipment, and facilities) for conducting JT&E's. Because of this constraint, the DDT&E arranges for a lead military service and a joint test director from one of the armed services to manage each test under the auspices of the DDT&E. Thus, the military services actually conduct the tests and then prepare the joint-test reports. There is no institutional memory in the DDT&E office. GAO found it difficult to find documents or persons who knew about past JT&E's. In response to a request from GAO, DDT&E staff members located documents and provided them to GAO and to the Defense Technical Information Center; by September 1983, the DDT&E office had assembled a microfiche library on JT&E. Funding for the JT&E program ranged from an estimated $3.8 million for fiscal year 1972 to a budgeted $50 million for fiscal year 1984. (pp. 8-10)

Planning for JT&E is done year to year; there is no long-range plan for selecting the subject of a test or for conducting JT&E's. The office of the Joint Chiefs of Staff (JCS)--which is responsible for joint military operations--the military services, and all other DOD offices are invited annually to submit nominations for JT&E. Through 1983, 30 JT&E's had been initiated, a majority (22 of 30) at the request of organizations within OSD. GAO found that only 2 of the 30 tests that have actually been initiated were submitted by the JCS. In commenting on the seemingly small number of requests for JT&E, JCS staff representatives told GAO

that the JCS has favored joint military exercises rather than testing and evaluation for obtaining information on joint military capabilities. (p. 11)

Recent legislation, the Department of Defense Authorization Act of 1984 (Public Law 98-94), provides for the establishment of a new test-and-evaluation organization in DOD that will report directly to the Secretary. As of December 1983, JT&E had not become a responsibility of this new office, and it is unclear how this recently enacted legislation will influence the organization or the management of the JT&E program. GAO believes that the findings and observations about the program presented in this report will be useful to DOD in its deliberations on how best to implement the new legislation. (pp. 7-8)

THE THREE CASE STUDIES

The IIR Maverick joint test

The Imaging Infrared (IIR) Maverick joint operational test and evaluation was undertaken in 1977 in order to assess the operational feasibility of attacking tanks and other ground vehicles with the IIR Maverick air-to-surface missile under battlefield conditions. From the test results, the joint test force (JTF)--the group that conducts JT&E's--concluded that the heat-seeking IIR Maverick has "impressive capabilities" and that it "should meet its operational requirements." (p. 30)

GAO found that, although the IIR Maverick test was completed in a very timely manner, its overall technical quality was poor. Contributing significantly to the shortfalls in quality were (1) the omission of important issues from the test design, (2) unrealistic test conditions, and (3) problems in analysis and reporting. (pp. 71-72)

For example, when providing close air support, pilots in combat normally find it necessary to distinguish between enemy and friendly ground forces so that they will not fire their missiles at friendly troops. In this test, however, no friendly ground forces were used. Thus, the task of distinguishing friendly forces from enemy forces was an omitted issue. (pp. 38-39)

Some of the test conditions were unrealistic. For example, four better-than-average pilots flew all the test missions in mostly excellent weather, in one small target area with the same cues for finding targets on every pass. This controlled test environment did not represent the range of battlefield conditions that would be encountered in actual combat while employing the IIR Maverick. (p. 71)

Analysis and reporting problems also were evident in this test. For example, the JTF did not establish and use formal criteria for deciding what test data to use in the analysis and what data to discard as flawed. This may have introduced bias into the conclusions about what the test results showed. In addition, the JTF did not fully report test results that indicated the weapon's technical and operational problems. Examples of such unreported problems include the fact that pilots were overloaded with work when missions were flown on totally cloudy days, problems of hitting targets when there was little difference between the temperature of the target and the temperature of its background, and problems the pilots might have in surviving enemy defenses. (pp. 71-72)

DOD used the IIR Maverick test results as support for its decision to develop the missile. GAO believes that the test results did not establish the operational efficacy of the missile system under the range of conditions that the system can be expected to encounter in combat. In addition, the test results can potentially be misused because of the JTF's incomplete and inaccurate reporting. Some of the useful test data were not reported. These unreported data revealed the difficulties of operating the missile system under certain battlefield conditions in the test scenario. (pp. 72-74)

The TASVAL joint test

The purpose of the 1979 JT&E called "Tactical Aircraft Effectiveness and Survivability in Antiarmor Operations" (TASVAL) was to address many of the complexities of conventional close air support in a central European conflict. This test was intended to provide data on how effectively close air support aircraft--specifically, Army helicopters and Air Force fixed-wing aircraft together--could assist ground forces. DOD anticipated that the results would be helpful

in determining what aircraft to buy and how to combine them in combat operations. (p. 75)

Although this test was an ambitious undertaking, with more than 100 players, GAO found that the quality of TASVAL was poor for several reasons. These include (1) the omission of important issues from the test design, (2) unrealistic test conditions, and (3) shortcomings in the analysis. (pp. 98-99)

For example, the time that aircraft pilots need to respond to a request for close air support is likely to influence combat effectiveness. Earlier joint testing had made information on response time available, but TASVAL did not take into account the differing response times of fixed-wing aircraft and helicopters; instead, it assumed that both aircraft would arrive at the battlefield at the same time. The important issue of variable response time was omitted from consideration. (pp. 81-82)

Although some of the test conditions were realistic, such as having both enemy and friendly forces on the ground, others were unrealistic. For example, the terrain and the climate of the California test site were unlike those of central Europe, so that the test results cannot be used to estimate combat capability in a central European conflict, which was the purpose of the test. Another example is that, although close air support is affected by battlefield visibility, factors that would normally affect battlefield visibility, such as smoke and fire, were not simulated. (pp. 82-83)

Analysis problems also affected the quality of TASVAL. The overall effectiveness of helicopters compared to that of fixed-wing aircraft could not be ascertained because certain features of the test were not considered in the analysis. For example, the numbers of flights at different times of day were not equivalent for the two types of aircraft. For another example, estimates of aircraft effectiveness and survivability were taken from mathematical models, the assumptions of which were not verified. (pp. 86-87)

The JTF reported specific conclusions for each test objective, but no overall conclusions were stated on the appropriate combination of

aircraft and the type of aircraft needed in close air support operations. The test was conducted in order to obtain the latter information. Army and Air Force officials reported, however, that TASVAL has been useful for its tactics, training, and testing lessons. (pp. 100-01)

From the shortcomings GAO identified in its review, GAO found that the TASVAL test results are of doubtful utility for estimating the effectiveness of close air support aircraft in central Europe. In addition, the test results were not timely. The Secretary of Defense requested that the test results be available by September 1978, but the JTF report was not published until May 1980. (p. 100)

The ACEVAL joint test

ACEVAL, or the Multiple Air-to-Air Combat JT&E, was conducted in 1977 in order to determine how the outcome of air combat is related to the numbers of friendly and enemy aircraft engaged under various conditions. It was the first major operational test that was highly instrumented for recording the data necessary for evaluating air-to-air combat performance. (p. 102)

As in the two previous cases, GAO believes that the quality of this test was poor. Contributing to the shortfalls in test quality were the same three problems: (1) the omission of important issues from the test design, (2) unrealistic test conditions, and (3) problems in reporting. (pp. 123-24) For example, the basic measure of air combat effectiveness is the degree to which overall mission objectives are accomplished. A mission objective might be to defend an airbase. In ACEVAL, the mission objective was omitted. The aircrews had no objective to attain in the test. This produced test results that are reflective more of the aircrews' gamesmanship than of what they might do in combat. (p. 110)

In ACEVAL, some of the test conditions were unrealistic. For example, aircrews in the test were allowed to fly from the test area into a "safe" area near the test range whenever conditions seemed threatening or unfavorable. Aircrews in real combat cannot always assume that they are moving into undefended or battle-free areas. Thus, the ACEVAL results may be biased in that they do not reflect the range

of conditions that aircrews might reasonably be expected to encounter. (p. 110)

In addition, various reporting problems were found in the ACEVAL report. For example, recommendations were made for hardware improvements in air-to-air missiles that have no basis in the test data. (pp. 114-15)

ACEVAL was proposed for completion in 1976, but this date slipped to 1979. The test was relevant to the requestor's need in that it provided empirical information on the outcome of air-to-air combat, but the test results precluded generalization to all air-to-air combat. Nevertheless, the data have been used for building mathematical models for predicting the outcomes of large-scale air-to-air combat. The JTF advised against this, and GAO concurs with the JTF's reservations. (pp. 124-26)

ACEVAL's results have been used more appropriately in further studies of air combat tactics and in improving testing. The most significant achievement of this test is its demonstration of the feasibility of instrumenting a highly complex test of air-to-air combat. (p. 126)

SUMMARY OF GAO'S RESPONSES TO SENATOR PRYOR'S QUESTIONS

GAO's review of 3 joint tests identified a number of shortcomings in the quality of JT&E and a number of areas in which management attention is needed. To respond to Senator Pryor's specific questions about the JT&E program, GAO drew upon its examination of the 3 JT&E's and its review of the management and organization of the JT&E program.

How independent is the DOD organization that is responsible for conducting JT&E from other DOD organizations that have vested interests in JT&E results? The office responsible for the joint testing and evaluation of DOD's weapon systems has not been independent of organizations with vested interests in JT&E results, since it reports to the same DOD office that is responsible for weapon-system development. In addition, joint tests have been managed, carried out, and partially funded by the individual services, which have vested interests in the results. The Congress, in recent legislation, has provided for the office of a Director of Operational Test

Tear Sheet

and Evaluation that is to report directly to the Secretary of Defense, but the JT&E program has not yet been placed under this new office. It is not yet clear how the legislation will affect the organization of the JT&E program or alter JT&E's dependence on the cooperation of the services for resources and capabilities. (pp. 127-28)

Who requests joint tests and evaluations and why? Most of the 13 JT&E's that were completed between 1972 and 1981 were requested by organizations within the Secretary's office. The JCS and the services have been infrequent requestors of joint tests. Without much involvement from the JCS, however, the primary purpose and the greatest expected usefulness of the JT&E program are jeopardized, since the information that it produces is intended to contribute to the decisionmaking of the JCS about joint military operations. In only 3 tests did two or more services perform their missions and roles in joint combat operations. The reasons for conducting JT&E's are multiple. Most of the 13 completed tests had more than one objective, but all focused primarily on the operational aspects of hardware, equipment, or testing techniques rather than on the ability of military personnel to use weapon systems jointly. (pp. 128-29)

Are JT&E problems defined to include critical operational issues? Factors important in judging operational effectiveness were omitted from each of the 3 JT&E's analyzed in depth by GAO. Omissions are sometimes not acknowledged in official JTF reports. Although JT&E is a complex process that obviously can never include all issues, those missing from the 3 joint tests were clearly integral, in GAO's opinion, to the main questions being addressed. Not acknowledging the tests' limitations harms both the quality and the usefulness of their results. (pp. 129-30)

Do the design and implementation of joint tests generate reliable and valid data about the operation of weapon systems, their limitations, and the concepts of their employment? GAO cannot make a judgment about the reliability of the test data from the 3 JT&E's--that is, about whether each test was controlled sufficiently for repeated testing under the same planned conditions to yield roughly the same results.

However, a valid test result accurately predicts combat performance. GAO believes that the validity of the 3 JT&E's can seriously be questioned. Unrealistic test conditions, together with problems of analysis and reporting, are the primary reasons why the validity of the results of the 3 JT&E's is questionable. (pp. 130-32)

Do the joint test-and-evaluation results that are reported accurately reflect the data that are collected? A major step in data analysis is that in which raw data that have been collected in the field are converted into test results (such as a percentage of targets hit). Although GAO found that JT&E reports of test results are usually accurate reflections of the data that were collected, GAO found that the data were often not qualified with respect to the tests' constraints. In some instances, the data were not given appropriate prominence in the test reports; in other instances, key data were omitted entirely from the reports. (pp. 132-33)

Do the conclusions and recommendations that are reported accurately reflect the test-and-evaluation results? Drawing conclusions and recommendations from test results is the last step in the data-analysis process. The conclusions and the recommendations in the joint test reports are not always supported by the test results. Some of the results provide no support for the conclusions that have been drawn, and some of the results lead to conclusions that differ from those stated in the JTF reports. For example, the IIR Maverick report contains the conclusion that, in general, the pilots detected targets easily, but the test results indicated that the pilots had difficulty under certain weather and battlefield conditions. In some instances, the JTF's recommendations propose modifications to missiles and electronic equipment, among other things, that were not tested. (p. 133)

Do the reports of the results address the concerns of the people who requested the JT&E's? GAO's analysis shows that the 3 case study JT&E reports sometimes addressed the concerns of the requestors and sometimes did not. Where the reports were not responsive to the concerns of a requestor, the problem could generally be traced to the omission of critical issues from a test design or to the establishment of test

conditions that were unlike a projected combat situation. (pp. 133-35)

How are JT&E results used? The requestors made little use of the 3 tests that GAO examined. However, the Congress rather than the requestor of the IIR Maverick test--the Defense Systems Acquisition Review Council--used part of the IIR Maverick test results, both reported and unreported, to deny the Air Force funds for producing the missile. The Air Force and the DDT&E, rather than the requestor of the ACEVAL test--DOD's program analysis and evaluation office--used the ACEVAL test results for computer model data in order to simulate air combat under different conditions using missiles with different capabilities. The ACEVAL results have also been cited on both sides of the debate about whether the U.S. weapons-acquisition strategy should emphasize quality or quantity. The appropriateness of either use seems questionable, given the test that was performed. (p. 136)

If the quality and usefulness of joint tests and evaluations are flawed, what are the possible reasons? The reasons for the threats to JT&E's quality and usefulness are complex and difficult to isolate. However, GAO believes that reasons for some of these threats may lie in the organizational features of the JT&E program. These include its organizational placement in the office of the Director for Defense Research and Engineering, its limited staff size, the failure to choose its staff members for their testing expertise, its limited budget, its dependence on the services for resources, and the absence of a strategic plan that sets priorities. (pp. 136-38)

RECOMMENDATIONS TO THE SECRETARY OF DEFENSE

GAO's finding that only 3 of the 13 JT&E's that were completed between 1972 and 1981 focused on joint operations indicates either that DOD does not perceive a need for JT&E information in making decisions about the combinations and structures of forces and the roles and missions of the services or else that DOD does perceive a need for JT&E data in addressing these issues and the JT&E program has not been responsive to this need. GAO recommends that the Secretary of Defense ascertain DOD's need for joint tests

that focus on the joint operations of the armed services. The JT&E program should be continued if the Secretary concludes that DOD has such a need. (p. 139)

If the Secretary of Defense determines that DOD does need the JT&E program, GAO recommends that the Secretary take the further steps that are necessary to (1) insure that priorities be established for conducting JT&E's, (2) endow the JT&E program with enough independence, permanence of expert staff, and control of resources to allow the program to conduct and report on joint tests and evaluations that both are high in quality and provide relevant information to their requestors and other users, and (3) require the JT&E program director to develop routine procedures that will insure that thorough records of test data, test results, and their use are maintained. (p. 139)

With regard to the implementation of these recommendations, GAO believes that the recently enacted legislation establishing an office of Operational Test and Evaluation in DOD may provide an opportunity to reduce the problems of JT&E's quality and usefulness that are shown in this report. If JT&E were to become a part of this unit--which, under the legislation, is to be independent of other DOD offices and agencies--then the organizational placement of the JT&E function might no longer pose a potential threat to test quality. However, JT&E's organizational independence is only a necessary condition; it is not in and of itself sufficient for achieving quality and usefulness, because it cannot automatically provide expertise, resources, user focus, or the coordination that is needed between service operations people and test analysts if JT&E's are to be sound. (p. 139)

AGENCY COMMENTS

GAO asked DOD to comment on a draft of this report. DOD elected not to comment.

Tear Sheet

Contents

		Page
DIGEST		i
CHAPTER		
1	OBJECTIVES, SCOPE, AND METHODOLOGY	1
2	JT&E's HISTORY, STRUCTURE, AND PROCEDURES	7
	History	7
	Structure	8
	Procedures	12
	Summary	13
3	ASSESSING THE QUALITY AND USEFULNESS OF JT&E THROUGH CASE STUDIES	15
	Examining the test process	15
	Assessing test quality	27
	Assessing the usefulness of test results	28
	Summary	28
4	THE IMAGING INFRARED (IIR) MAVERICK JT&E	30
	The context of the IIR Maverick test	32
	The test objectives and design	33
	The quality of the test results	36
	Summary of quality	71
	The usefulness of the test results	72
5	THE JOINT TACTICAL AIRCRAFT EFFECTIVENESS AND SURVIVABILITY IN CLOSE AIR SUPPORT ANTI-ARMOR OPERATIONS (TASVAL) TEST	75
	The context of the TASVAL test	76
	The test objectives and design	77
	The quality of the test results	80
	Summary of quality	98
	The usefulness of the test results	100
6	THE MULTIPLE AIR-TO-AIR COMBAT (ACEVAL) JT&E	102
	The context of the ACEVAL test	103
	The test objectives and design	104
	The quality of the test results	107
	Summary of quality	123
	The usefulness of the test results	124
7	CONCLUSIONS AND RECOMMENDATIONS	127
	Conclusions	127
	Recommendations to the Secretary of Defense	139

		Page
APPENDIX		
I	Congressional request letter	141
II	Bibliography	143
III	Supplementary data on DOD's JT&E activity 1972-83	153
IV	Technical data for chapter 4 on IIR Maverick	162
V	Technical data for chapter 5 on TASVAL	178
VI	Technical data for chapter 6 on ACEVAL	197
FIGURE		
1	The four categories of DOD testing and evaluation	3
2	The steps of our review of the quality and usefulness of JT&E	4
3	The illustrative characteristics of the case study tests	5
4	A profile of DDT&E's fiscal year budgets 1972-84	9
5	The division of management levels for a typical joint test force	12
6	The three phases of our JT&E case study analysis	15
7	The seven steps of the JT&E process	15
8	Step 1: Understanding the context of the test	16
9	Step 2: Defining the test objectives	17
10	The elements of battle and three models	18
11	The battle as the interaction of factors in the employment of a weapon system	19
12	Step 3: Planning the test	20
13	Step 4: Implementing the test	23
14	Step 5: Analyzing the data	24

FIGURE		Page
15	Step 6: Reporting the results	26
16	Step 7: Using the results	27
17	The IIR Maverick test objectives	31
18	Design matrix for the IIR Maverick JT&E	34
19	The major variables considered in the IIR Maverick JT&E	34
20	The IIR Maverick JT&E scenarios	35
21	The process of employing the IIR Maverick	35
22	Threats to test quality: The transition objective	38
23	A-7 pilot learning: Target-area acquisition range	44
24	Threats to test quality: The valid target objective	48
25	The probability of a pilot's acquiring a valid target with the IIR Maverick	52
26	Threats to test quality: The cueing objective	55
27	Threats to test quality: The survivability objective	59
28	Threats to test quality: The single-seat aircraft objective	62
29	Threats to test quality: The counter-measures objective	66
30	Threats to test quality: The thermal character objective	68
31	Summary of the quality of IIR Maverick test results	71
32	The TASVAL final test objectives	77
33	Design matrix for the TASVAL JT&E	78

FIGURE		Page
34	The major variables considered in the TASVAL JT&E	78
35	The process of providing close air support in TASVAL	79
36	Threats to test quality: The effectiveness objective	81
37	Threats to test quality: The attrition objective	90
38	Threats to test quality: The synergism objective	96
39	Summary of the quality of TASVAL results	99
40	The ACEVAL test objectives	104
41	Design matrix for the ACEVAL JT&E	105
42	The major variables considered in the ACEVAL JT&E	105
43	The process of air-to-air combat in ACEVAL	106
44	Loss rates for force ratios for trials with two friendly aircraft	108
45	Threats to test quality: The aircraft numbers objective	109
46	The effect of the advantage of having ground control intercept information	116
47	Threats to test quality: The ground control intercept objective	117
48	The relation between force ratio and exchange ratio for two aircraft types	119
49	Threats to test quality: The aircraft type objective	120
50	Summary of the quality of ACEVAL results	124
51	Summary of the possible organization-based threats to the JT&E process that can lead to low quality and low usefulness	137

ABBREVIATIONS

ACEVAL	Multiple Air-to-Air Combat Evaluation
AFTEC	U.S. Air Force Test and Evaluation Center
AHT	Attack helicopter team
AIMVAL	Air Intercept Missile Evaluation
CAS	Close air support
DDR&E	Director for Defense for Research and Engineering
DDT&E	Director for Defense Test and Evaluation
DOD	U.S. Department of Defense
DSARC	Defense Systems Acquisition Review Council
DT&E	Developmental testing and evaluation
EW/CAS	Electronic Warfare in Close Air Support
GAO	U.S. General Accounting Office
GCI	Ground control intercept
IDA	Institute for Defense Analyses
IIR	Imaging infrared
JCS	Joint Chiefs of Staff
JOT&E	Joint operational testing and evaluation
JT&E	Joint testing and evaluation
JTF	Joint test force
OSD	Office of the Secretary of Defense
OT&E	Operational testing and evaluation
PPI	Preplanned interdiction
SPC	System Planning Corporation
TASVAL	Tactical Aircraft Effectiveness and Survivability in Close Air Support Anti-Armor Operations
WSEG/IDA	Weapon System Evaluation Group and Institute for Defense Analyses

CHAPTER 1

OBJECTIVES, SCOPE, AND METHODOLOGY

In most military combat situations, the U.S. Department of Defense (DOD) combines the operations of critical military missions in a way that transcends the boundaries and responsibilities of the individual armed services. Consequently, how the Air Force, Army, and Navy interact during combat is extremely important. What do we know about their ability to perform joint missions or to conduct combined operations? In 1970, the Blue Ribbon Defense Panel determined that "there is no effective method for conducting OT&E [operational testing and evaluation] which cuts across Service lines . . ." (II.A.3, p. 90).[1] Basing its decision on this concern, DOD made joint testing and evaluation (JT&E) a formal activity in 1971.

What has happened in the 13 years since then? In 1978, DOD officials recognized that the JT&E program lacked discipline, an overall structure, and a clear concept of the needs of the services, the Joint Chiefs of Staff (JCS), and the Office of the Secretary of Defense (OSD). They requested a study to find solutions to these problems, and the result was the creation in 1979 of a management framework for JT&E. Today, the question remains: Has productive JT&E been accomplished?

The Honorable David Pryor of the U.S. Senate asked us to review the quality, limitations, and usefulness of DOD's joint testing and evaluation--that is, tests that are supposed to examine issues transcending the individual services, especially those assessing the ability of the United States to perform military missions in a joint environment.[2] After discussion with him, we posed the following questions:

--Who requests joint tests and evaluations and why?

--How independent is the DOD organization that is responsible for conducting JT&E?

--Do the definitions of JT&E problems include critical operational issues?

---

[1]The bibliographic data for the source of the quotation (and all quotations in this report) are in appendix II. Here, "II.A.3" means that this quotation's source is the volume of the 1970 report of the Blue Ribbon Defense Panel that is listed in appendix II, section A, item 3. (Appendix I contains the letter from Senator Pryor asking us to conduct this review.)

[2]We use "JT&E," "joint testing and evaluation," "joint test," "test," and other such expressions interchangeably, except where the meaning or context requires otherwise.

--Do the design and implementation of joint tests generate valid and reliable data about weapon systems' operations and limitations and the concepts for their use?

--Do the results, conclusions, and recommendations that are reported for JT&E accurately reflect the test data?

--Do JT&E reports address the concerns of the people who requested them?

--How are JT&E results used?

--If the quality and usefulness of JT&E are flawed, what are the possible reasons?

Military testing and evaluation make up a structured investigation whose purpose is to obtain, verify, and supply data for making some assessment or judgment. DOD's Directive 5000.3, entitled "Test and Evaluation," establishes policy and designates responsibility for the four categories of testing and evaluation that we list in figure 1. In contrast to other military testing, JT&E is intended to transcend service boundaries and to help in determining the most effective combination of forces, force structures, and procurement alternatives; in establishing the requirements for improving equipment or systems; and in developing the mission and activity of the JCS and the individual services. As the figure shows, the responsibility for insuring that JT&E is productive belongs to DOD's Director for Defense Test and Evaluation (DDT&E).

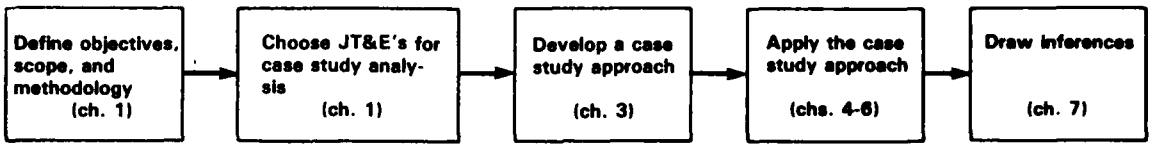
In earlier reports, we have described DOD's test-and-evaluation enterprise (II.A.14, 16-19, 21, 22, 24, 28-31), assessed the operation of the testing organizations of individual services (II.A.15, 16, 23, 26), and examined the results of tests conducted during the acquisition of specific weapon systems (II.A.17, 20, 25, 27, 28). In answering Senator Pryor's questions for this report, we established two main objectives: to assess the quality and usefulness of JT&E results and to find out the systemic reasons for whatever limits there are to their quality and usefulness. To meet these objectives, we adopted two others: to describe what JT&E is and how it is organized and to determine how well DOD's management of JT&E has led to productive joint operational tests. The steps of our review are outlined in figure 2 on page 4.

In order to describe the JT&E program, we reviewed its history, organization, policy guidance, and operating procedures and the management of the office of the DDT&E. Our understanding and judgments in this segment of our review are based on the interviews we held with past and present DOD officials and contractors who have been or are involved with JT&E and on the analysis we made of the DOD documents we cite in the bibliography in appendix II. Among these, we relied heavily on the 1970 study by

Figure 1
The Four Categories of DOD Testing and Evaluation

Category	Purpose	Result	Responsibility
Developmental (DT&E)	To assist the engineering and development process and verify the attainment of technical performance specifications and objectives	Data used for selecting system concepts, identifying the technical approach, correcting system design problems, and developing, deciding to accept, and introducing system changes	Services' materiel development agencies
Operational (OT&E): Initial (IOT&E) and Follow-on (FOT&E)	To estimate a system's operational effectiveness and suitability and identify needed modifications and to provide information on tactics, doctrine, organization, and personnel requirements	Data used for preproduction alternative, modification, and production decisions (IOT&E); data used to insure that initial production items meet operational effectiveness and suitability thresholds and to evaluate system, personnel, and logistic changes (FOT&E)	Services' independent OT&E organizations
Multiservice (MT&E)	To provide DT&E or OT&E data for systems being acquired jointly or operated jointly with equipment of another DOD component	Same as DT&E and OT&E	DT&E and OT&E organizations in two or more DOD components
Joint (JT&E)	Primarily to examine the ability of developmental and deployed systems to perform their intended missions in a joint environment; secondarily to provide information for technical concepts evaluation, system requirements or improvements, systems interoperability, force structure planning, developing or improving testing methodologies, and doctrine, tactics, and operational procedures for joint operations	Data used with other information as a basis for systems acquisition, mission area analysis, program decisions on related forces, force issues, and operational procedures and doctrine	Required and initiated by the Director for Defense Test and Evaluation, who delegates responsibility for each JT&E to a specific service; participating services provide forces and materiel

Figure 2
The Steps of Our Review of the Quality and Usefulness of JT&E



the Blue Ribbon Defense Panel and the 1979 study by the BDM Corporation (II.A.1, 3-5). Our overall description of JT&E is in chapter 2.

Since our goal in determining the quality and usefulness of test results was to trace any limitations we discovered back to their origins in the test process, we chose the case study method as one that would give us the most detailed information about individual tests. (Our definitions of "quality" and "usefulness" are in chapter 3.) The 3 JT&E's we chose for analysis were among the 13 major tests that had been completed by January 1981. In the order that we discuss them, the 3 JT&E's are the Imaging Infrared (IIR) Maverick test, the Joint Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations test (TASVAL), and the Multiple Air-to-Air Combat Evaluation test (ACEVAL).

We used the following criteria for selecting these tests. First, we considered only completed tests, so that we would be able to review the entire test process, from the context in which a test is begun to the use of the results. Second, given the 13 tests completed by January 1981, we chose from among the more recently completed tests so that we would be able to interview the test managers and participants while their experience was still within easy recall. Third, we looked for tests that would illustrate the several types of JT&E, given that the purpose, requestors, and participation of the armed services differ from test to test. In figure 3, we summarize the illustrative characteristics of the three tests we selected.

For each test, we performed the following activities. (1) We reviewed the reports on it, including feasibility studies, designs and plans, and final reports prepared by the joint test force, the independent contractors, and the individual services. (2) We reviewed memorandums, original data-collection forms, and other pertinent information from the files of the DDT&E, the test's managers, and the contractors. (3) We reviewed legislation, reports, and articles that used the test data. (4) We held semistructured interviews with individuals who had been involved with the test, including the DDT&E test monitors, the independent contractors, representatives of the joint test force and the services, and the test's participants. (5) We reviewed other tests on similar issues. (6) For ACEVAL and TASVAL, we visited the test

Figure 3
The Illustrative Characteristics of the Case Study Tests

Test	Requestor	Intended use	Services involved	Test scenario	Performance assessed	JTF report
IIR Maverick	Deputy Secretary of Defense	Procurement	Air Force (responsible for implementation), Army, Navy	Air Force: close air support and preplanned interdiction	Friendly force	July 1977
TASVAL	Secretary of Defense	Air Force and Army acquisition, force structure, and force combinations	Army (responsible for implementation), Air Force, Marine Corps	Army: close air support, Air Force: close air support, Army and Air Force: joint close air support	Friendly and enemy forces	May 1980
ACEVAL	Program Analysis and Evaluation, Office of Secretary of Defense	Force structure and test procedures	Navy (responsible for implementation), Air Force	Navy: air to air Air Force: air to air	Friendly and enemy forces	February 1978

sites and looked at the instrumentation; for the IIR Maverick test, we reviewed subsequent operational testing and observed two test missions at Ft. Drum, New York, in 1982. Our review was performed in accordance with generally accepted government audit standards. Although we asked DOD to comment on a draft of this report, DOD elected not to provide either oral or written comments.

Within the limits of our three cases, we sought to draw some inferences about the way JT&E is conducted. We focused on the systemic strengths and weaknesses of the tests, concentrating on the quality of the test-and-evaluation process as revealed in the tests we analyzed.

In structuring this report, we have placed our overall description of JT&E in chapter 2, in which we summarize its history and the organization, policy guidance, and managing and operating procedures that are generally used in conducting joint tests. We supplement this with appendixes I-III, which contain Senator Pryor's request letter, a reference list for the documentary sources we cite, and background information on the JT&E program.

In chapter 3, we explain our method of assessing quality and usefulness and give a step-by-step description of the activities and decisions entailed in the test process that the joint test forces use in conducting joint tests. Understanding the method we followed and the seven steps of the test process are prerequisites for interpreting our analysis of the three tests in chapters 4, 5, and 6. Appendixes IV, V, and VI add detail to our analyses and

inferences, providing much of the technical data that form the basis of our statements.

In chapter 7, we state our findings about the individual tests and about JT&E as a whole. The recommendations we present in chapter 7 are based on these findings.

CHAPTER 2

JT&E'S HISTORY, STRUCTURE, AND PROCEDURES

Since DOD's Office of the Director for Defense Test and Evaluation is responsible for initiating and coordinating productive joint testing, a discussion of the history, structure, and procedures of JT&E necessarily focuses on that office. Our information derives primarily from two reports. One is the 1970 report of the Blue Ribbon Defense Panel, which was appointed in 1969 by the President and Secretary of Defense to study the organization, structure, and operation of the Department of Defense. Its recommendations led to the establishment of the JT&E program. The other is the 1979 report by the BDM Corporation on JT&E's progress and the ways in which it could be improved. However, we also reviewed other pertinent documents and interviewed DOD officials and contractors (as we noted in chapter 1).

HISTORY

In the 1960's, it was recognized that DOD lacked an effective method of insuring that decisionmakers had information from operational tests and evaluations about joint tactics and operating procedures in combat situations. In 1968, "the Deputy Secretary of Defense requested the JCS to consider the establishment of a small Joint Test and Evaluation Agency" (II.A.3, p. 89). The response of the JCS was that existing DOD organizations made this unnecessary. In 1970, the Blue Ribbon Defense Panel recommended that DOD create a defense test agency, with a civilian director, to conduct tests and evaluations that would help detect deficiencies, predict combat capability, and support decisions about systems, equipment, and the composition of forces with adequate information about how the military services interact when they combine their operations.

Although DOD did not create a separate test agency, in 1971 the position of Deputy Director for Test and Evaluation was established within DOD's office of the Director for Defense Research and Engineering. In 1977, the position was changed to Director for Defense Test and Evaluation (DDT&E) and, in 1979, the location was changed to the office of the Under Secretary for Defense Research and Engineering. Except for the period between April 1978 and December 1979, when OSD's Program Analysis and Evaluation office was given direct responsibility for operational testing and evaluation, the DDT&E has maintained broad responsibility for test-and-evaluation matters, having been explicitly directed to initiate, coordinate, and insure that the military services conduct productive, objective, and timely operational tests and evaluations.

A law that became effective on November 1, 1983, provides for a civilian Director of Operational Test and Evaluation in the Department of Defense, to be appointed by the President with the

advice and consent of the Senate. This director will be OSD's principal advisor on operational test-and-evaluation matters and DOD's senior operational test-and-evaluation official. Although the relationship of this new office to the DDT&E is presently unclear, some of its functions will be similar to those of the DDT&E. (The law is the Department of Defense Authorization Act of 1984 (Public Law 98-94), the pertinent passages of which are reprinted in appendix III.)

STRUCTURE

Leadership and personnel

The DDT&E position has been filled by civilians who were recently retired from the military. The first two DDT&E's had been retired for less than one month when they were selected. The third, the DDT&E at the time of our review, was chosen before he retired from military service.

Few staff members are assigned to JT&E functions--approximately seven people in the office of the DDT&E spend less than 30 percent of their time on JT&E. They are, with one exception, career military officers on rotation for two or three years from their services. The rotations mean that knowledge about JT&E in the office of the DDT&E is not cumulative.

From an interview with a DDT&E official, we learned that the recruitment of staff for the DDT&E does not emphasize training and experience in operational testing and evaluation. Before the office of the DDT&E was formed, the Weapon System Evaluation Group, which comprised 50 senior officers from all the services, had become involved in several studies and tests on joint operations for the JCS and the Under Secretary for Research and Engineering. It was suggested that the OT&E capability of this group be expanded in support of the JT&E staff, but it was disbanded in 1976, and no other DOD office has replaced it.

The Defense Technical Advisory Board was established in 1980 and is available to help the DDT&E with technical issues when JT&E's are being nominated, selected, conducted, and evaluated. This Board consists of 12 civilian scientists who are employed by and have full-time responsibilities within the services. They meet periodically--usually once a year or at the call of the DDT&E--and support the DDT&E strictly as advisors.

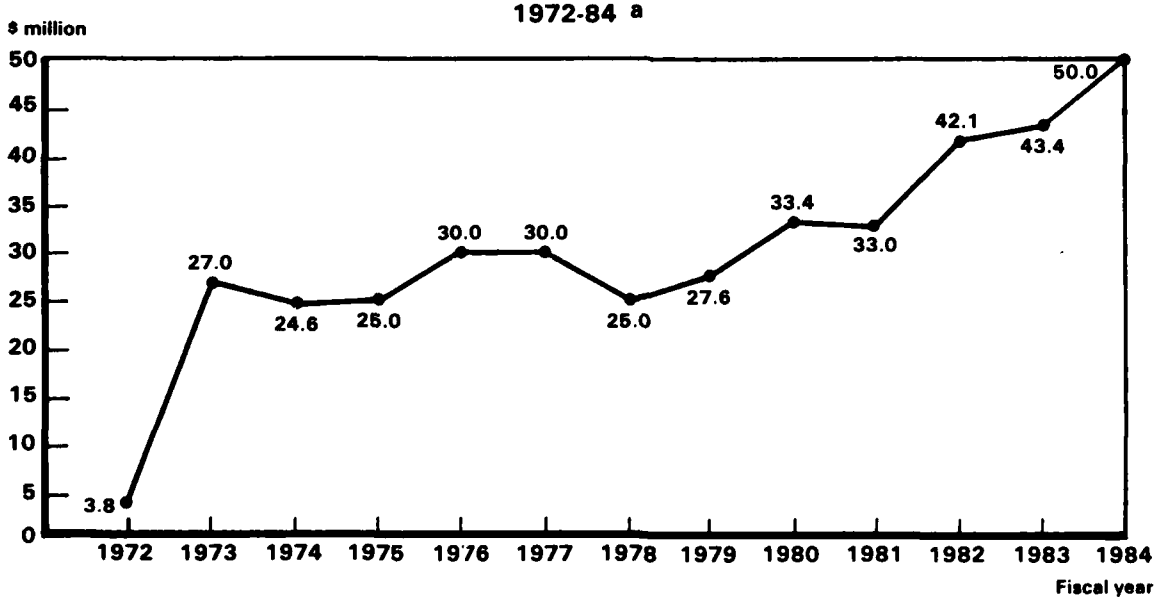
Thus, we found no institutional memory in the DDT&E office. It was difficult to find documents or even persons who knew about past JT&E's. When we first requested JT&E documents, in the spring of 1981, they were not available in the DDT&E's office or at the Defense Technical Information Center. In response to our request, the DDT&E staff located documents and gave them to us and to the Center. By September 1983, the DDT&E's staff had assembled a microfiche library on JT&E (II.A.2).

Funding and other resources

The DDT&E does not have either total or the preponderant control over resources for JT&E. Costs for joint testing and evaluation are covered by two primary sources--a separate OSD appropriation for DDT&E (program element 65804D) and funds from the services. In figure 4, we show DDT&E's annual funding since 1972. By far the greatest amount of the DDT&E's budget has been spent on directing and supervising specific JT&E's and on costs that are unique to JT&E (developing, procuring, installing, and operating special instrumentation, for example). About 8 percent has paid for feasibility studies, facilities, instrumentation, and other items that were not pertinent to specific joint tests. The DDT&E's limited resources force reliance on contractors for feasibility studies, test designs, data analyses, and assessments of test results, all of which are largely performed by such organizations as the Institute for Defense Analyses, the System Planning Corporation, and the BDM Corporation.

Since the funds from OSD's separate appropriation for DDT&E do not cover the costs of JT&E test sites, personnel, and equipment, the services must provide these resources. Until recently, the services did not have accounting systems set up in a way that would reflect their costs, however. The BDM Corporation, in drawing conclusions for its 1979 report, could only indicate estimates of JT&E costs covered by the services, suggesting that they have been approximately equal in the aggregate to the OSD appropriation while undoubtedly varying from test to test.

Figure 4
A Profile of DDT&E's Fiscal Year Budgets 1972-84 [a]



[a] Figures have not been adjusted for inflation. Fiscal 1976 includes $5 million for the transition year.

According to the DDT&E's January 21, 1983, budget summary, 30 joint tests have been initiated since 1972. (A list of these tests is in appendix III.) It was estimated at the end of fiscal year 1983 that 24 would then be completed or terminated for other reasons.

Since the office of the DDT&E does not have the money to reimburse the services for critical test resources, it must rely on their cooperation and good will. It has found, however, that some resources, such as personnel and equipment, have been difficult to obtain. The services contend that having a large amount of resources tied up in a joint test hinders their training and operational readiness. They add that it is difficult to include the OSD's nominations for tests in their long budgetary projections. Test nominations include an estimate, but it is difficult to project costs before test planning is well under way, because sites, force configurations, instrumentation, and duration vary greatly from test to test.

Having no control over any test facility, the DDT&E depends on the services for test sites. Moreover, the DDT&E has no overall plan for addressing JT&E issues, so that test-related equipment that the DDT&E could command for repeated use in a number of different tests has never been developed. Most of the equipment that the DDT&E procures for each JT&E is given to the services at the conclusion of the test.

The office of the DDT&E is, however, attempting to maintain greater control over the equipment that it owns. The less costly items are grouped together and, like the more costly items, are set forth in a memorandum of understanding by which the services will manage the equipment. The DDT&E keeps first priority for the use of the instruments that are developed for JT&E programs and reserves the right to approve major modifications to them. Whether all this means that the equipment the DDT&E buys for one JT&E will be used in subsequent JT&E's remains to be seen.

Management and accountability

The DDT&E has ultimate responsibility for insuring that productive JT&E's are conducted but arranges for a lead service and a joint test director to manage each test. This makes for discontinuity in the individual and overall management of JT&E.

The DDT&E

After the DDT&E receives nominations for JT&E's, the Joint Test and Evaluation Planning Committee reviews them and recommends to the Senior Advisory Council specific tests to be conducted and their priority. The Council (composed of officers from the services and JCS) decides which ones it believes should be conducted, and a private company on contract to DDT&E studies their feasibility. From the results of the feasibility studies,

the DDT&E makes a final decision about whether or not to go ahead with each test. This decision has a great deal to do with what the services want.

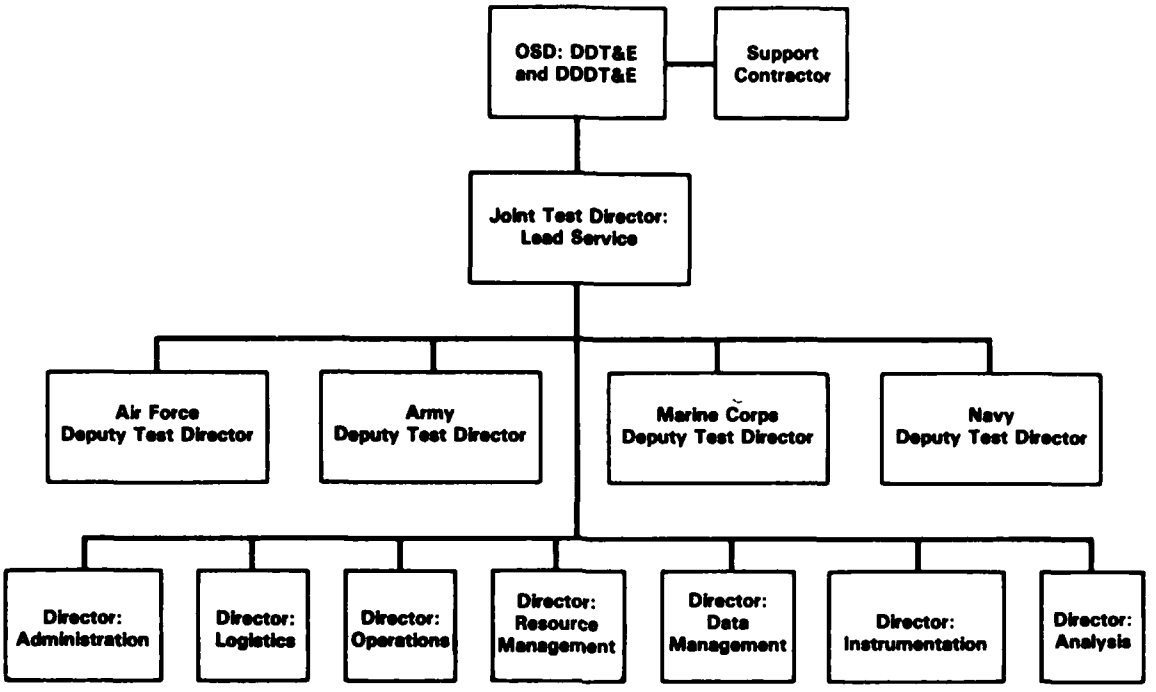
Moreover, the JCS has been noticeably absent among the major requestors of joint tests. The JCS is DOD's main proponent for joint procedures and the interoperability of deployed forces. DOD Directive 5000.3, noting that the JCS requires JT&E information on doctrine, tactics, and operational procedures, specifically directs it to coordinate the annual nominations for JT&E that it makes with those from the services and the commanders-in-chief of the unified and specified commands (responsible for determining and implementing joint doctrine and combined arms concepts). In addition, the JT&E procedures manual specifies that all joint tests that concern issues related to joint doctrine or missions must be submitted to the JCS for concurrence and coordination. However, the JCS has shown very little interest in JT&E, as evidenced by its nomination of only 2 of the 30 tests actually initiated. Nor has the JCS actively sought nominations for JT&E's from the services. The JCS has stated its belief that the experience and free play of field exercises provide more valuable and timely information than JT&E's quantitative data. In a November 1982 memorandum to the Director of the JCS, the DDT&E solicited help with reviewing JT&E nominations, but the JCS did not respond.

Concerned that the JCS and the services were not involved in JT&E, the DDT&E began using new procedures in 1981 in the hopes of increasing their participation. They are now represented on the Joint Test and Evaluation Planning Committee and the Senior Advisory Council and do participate in the more formal processes of nomination and selection, but to what extent we were not able to determine. A September 1983 planning committee meeting resulted in only one nomination for a joint test, and the requestor was DDT&E. The JCS is currently forming a task group with service participants to review the JT&E program and develop recommendations to improve its operation. The Air Force and the Army have established offices within their operational test-and-evaluation organizations to coordinate joint testing issues with the office of the DDT&E.

The joint test director

Once a JT&E's objectives and design have been formulated, the DDT&E has limited involvement with it, so that managerial continuity is interrupted. Each test has its own set of managers, who report through a joint test director to the DDT&E. In the organization of a typical joint test force, the joint test director oversees the test, but directors appointed from each of the participating services manage the allocation and use of their respective resources. In TASVAL and ACEVAL, for example, the aircrews reported to the Air Force and the ground forces reported to the Army, and each service had full control over its own personnel ratings.

Figure 5
The Division of Management Levels for a Typical Joint Test Force



In the past, JT&E's did not have full-time directors. The joint test director for IIR Maverick was concurrently the commander of the Air Force Test and Evaluation Center. The joint test director for TASVAL was also the commander for the U.S. Army Combat Developments Experimentation Command. For more recent tests, the DDT&E has appointed full-time directors from the services who are free from other responsibilities while they are test directors. Figure 5 shows the typical division of management levels.

No single entity is solely accountable for the formulation and execution of JT&E from start to finish. The DDT&E provides somewhat limited guidance on test implementation, the joint test directors find it difficult to take complete control of the test settings, equipment, and participants, and the services tend to vest more interest in their individual objectives than in those of the joint effort.

PROCEDURES

The official DOD directive on military testing and evaluation (Directive 5000.3) establishes policy, designates overall responsibilities, and sets forth pertinent definitions. In 1980, the BDM Corporation prepared a JT&E procedures manual for the DDT&E, who was trying to give the program more definitive guidance, give

structure to the nomination and selection processes, and spell out in greater detail how the JCS and the services should be involved in conceptualization and design. The manual's "new or baseline JT&E architecture," as DOD calls it, would clarify the responsibilities and authorities of the DDT&E, the JCS, the services, and the joint test directors, but it would leave the organizational structure of JT&E essentially unchanged. The manual has not been made official, because the Navy disagrees with the JT&E process that is proposed in it (see appendix III).

It is too early to determine whether following the manual will improve JT&E's results. The DDT&E began following the procedures unofficially for the tests nominated in 1979 for which budget authority began in fiscal 1982, but none of those tests has been completed. The solicitation of JT&E's has been standardized, but the nominations are still made ad hoc. We found no evidence of an overall agenda or a strategic plan for addressing JT&E issues.

The BDM Corporation has observed that joint tests have "little relationship to one another or an overall JT&E program. There are also many systems or concepts which have not been tested in a joint setting" (II.A.1, p. III-2). The tests cannot be linked with the development programs within OSD or the services. Without an overall plan that states priorities, there is no assurance that the many issues, concepts, and systems that have not been tested in joint settings will be addressed or that the most important issues will have first priority.

Among the 13 tests that had been completed at the time of our review, for example, most had more than one objective although the focus was generally on gathering information. Three were intended to provide data for weapon-system acquisition decisions, 4 were to establish whether the hardware or system design requirements or the operational capabilities of deployed or developmental systems could be met, 2 were to determine the utility of the procedural or technical concepts for existing or developmental weapon systems, and 4 were to evaluate techniques for improving testing methodology. As for joint participation, the Air Force was chartered to participate in 12 of the 13, the Army in 9, the Navy in 10, and the Marine Corps in 3. The JCS, with the U.S. Readiness Command, participated in only one. Although two or more services participated in all 13 tests, they rarely combined their operations. The services actually performed jointly in only 3, one of which was TASVAL.

SUMMARY

DOD's JT&E program has concentrated very little in its 13-year history on the ability to perform joint military missions. The organizations in DOD with the greatest responsibility for combined military operations--the Joint Chiefs of Staff and the armed services--do not view joint test-and-evaluation activity as a significant source of information. Even though two or more

services have participated in joint tests, very few of the completed JT&E's have involved joint operations. The DDT&E has developed no overall strategy or plan to insure that the JT&E program can or will address joint issues.

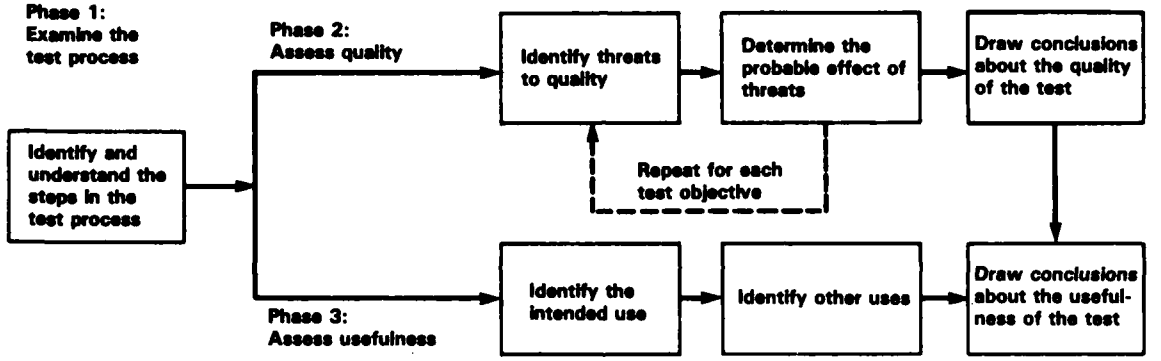
While productive JT&E is the function of a central organization within OSD, the tests are managed, carried out, and even partially funded by the separate services. The DDT&E does not have power, authority, or continuity in the management of JT&E. The DDT&E is affiliated with the Under Secretary of Defense for Research and Engineering and must rely for resources on the cooperation of the services. The unofficial procedures that the DDT&E has been following since 1980 standardize the process of initiating tests and clarify the responsibilities of the several levels of test management, but they do not make the organizational structure of the JT&E program less dependent on the armed services or the developers of DOD's weapon systems.

CHAPTER 3

ASSESSING THE QUALITY AND USEFULNESS OF JT&E THROUGH CASE STUDIES

In this chapter, we explain the method we used to examine the IIR Maverick, TASVAL, and ACEVAL JT&E's that we discuss in chapters 4, 5, and 6. In all three cases, we were looking for well-formulated questions about the ability of the armed services to combine military operations that led to tests that were designed well, implemented properly, analyzed appropriately, and reported scrupulously. As we show in figure 6, we examined the test process and then we assessed the quality of the three tests' results as well as their usefulness.

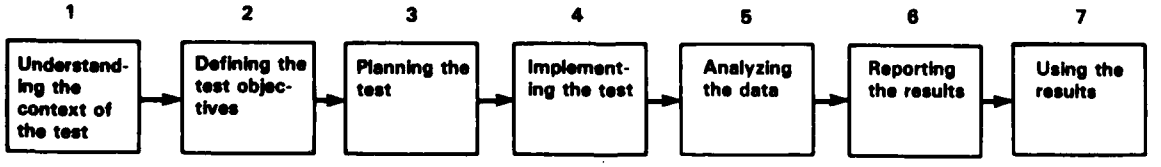
Figure 6
The Three Phases of Our JT&E Case Study Analysis



EXAMINING THE TEST PROCESS

Understanding the JT&E process as a whole is a prerequisite for analyzing any individual test. We found it helpful to look at the test-and-evaluation procedure of the joint test forces as a seven-step process, which we display in figure 7. The steps provide a conceptual framework for analysis that is grounded in our review of DOD's test-and-evaluation literature, although we know of no one military document that sets forth these seven steps comprehensively in a statement of doctrine.

Figure 7
The Seven Steps of the JT&E Process



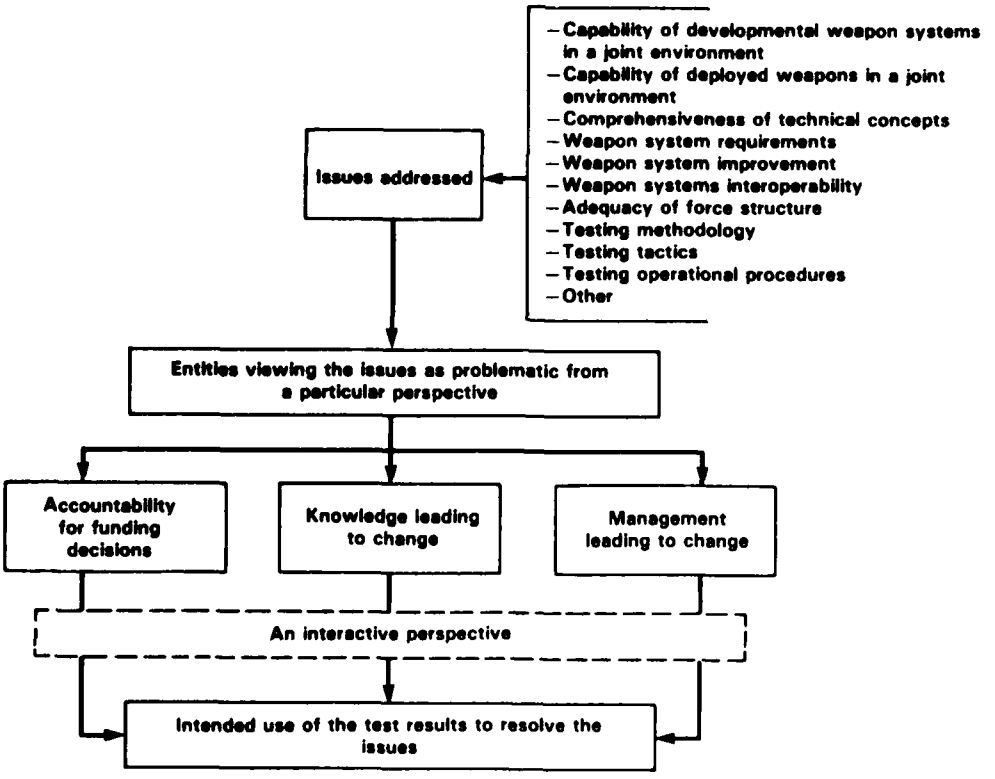
Step 1: Understanding the context

Understanding the context is the first step in the test process. The context of an operational test is made up of issues related to military performance. Can the armed services perform specific missions? What tactics and operating procedures should they use? How will organizations, functions, and persons be affected by the answers to these questions? What is the reason for expecting a specific JT&E to produce answers, and how are they likely to be used? JT&E usually addresses such issues, as we show in figure 8.

When the Congress, OSD, JCS, or any one or more of the individual services have questions about a particular aspect of military performance, they may view an issue as problematic. They may be accountable for funding decisions, they may need knowledge to make force-planning, deployment, or other decisions, or they may be engaged in management activity related to the issue. Their concern may come from some combination of these perspectives.

When JT&E issues are looked at from an accountability perspective, it is usually to seek information about the best possible use of resources. This happens when the Congress holds

Figure 8
Step 1: Understanding the Context of the Test



DOD responsible for its funding decisions. Since the services y not want the agencies and individuals who control their finan- al resources to obtain test evidence that calls their operation- effectiveness into question, an inquiry based on accountability n be threatening.

The purpose of looking at issues from the knowledge perspec- ve is generally to predict combat performance without reference immediate decisions for any one weapon system or tactics for s use. Substantive knowledge and appropriate change are the als, and they are sought by the Congress and, in particular, by e services. The search for general knowledge may become threat- ning, however, when it is tied to specific decisions, as in plan- ng how to structure and combine the armed forces.

Tests can be perceived as management tools for improving the verall efficiency of military operations. This seldom happens ithin the individual services, but OSD often has questions about oint military operations. Test results are more likely to be eld confidential when they are sought solely for management pur- oses than when their purposes are accountability or knowledge-- hen change occurs, there is less need for publicizing it.

Test issues may be viewed from more than one perspective. t is difficult to look at a problem purely for purposes of know- edge, for example, when the managers of a specific mission or eapon system are being held accountable for a funding decision. imilarly, the resolution of a management issue may contribute nowledge whose utility goes beyond an immediate management need.

tep 2: Defining the test objectives

The challenge at step 2 is to understand the elements of a omplex question and turn them successfully into research objec- ives. As we show in figure 9, this usually means in JT&E that

Figure 9
Step 2: Defining the Test Objectives

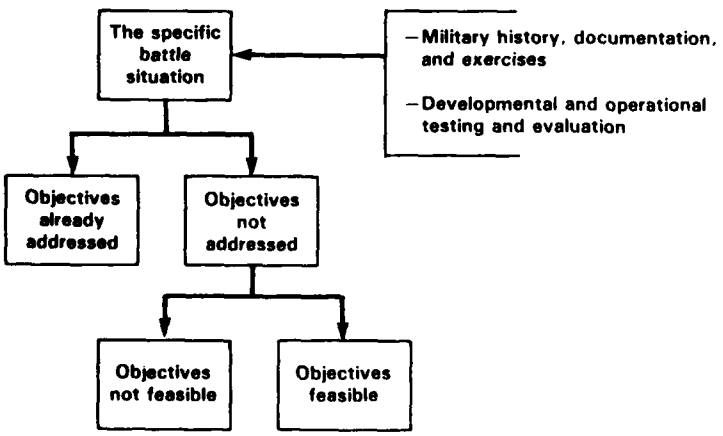


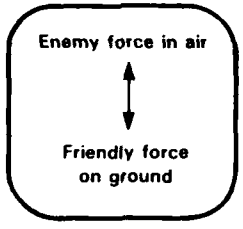
Figure 10
The Elements of Battle and Three Models

Levels of support	Pairings of forces	Concepts of interaction
Air Ground Sea	One vs. one One vs. two Two vs. two Three vs. one etc.	Interaction = two sides attacking each other (◄►) Noninteraction = one side attacking another (►) Negative interaction = one side attacking itself (►)
Two simple models	A more complex model	A very complex model

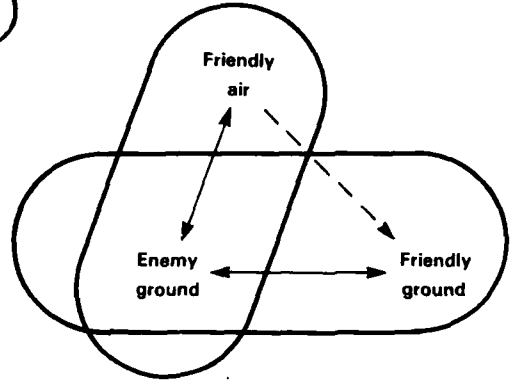
a. One-paired fight on one level, interactive:

Enemy force in air ◄► Friendly force in air

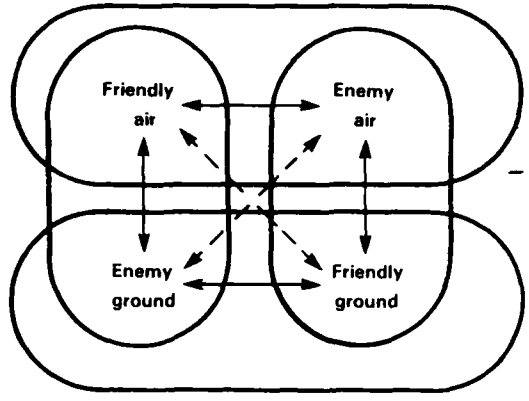
b. One-paired fight on two levels, interactive:



Two-paired fight on two levels, interactive and negative interaction:



Four-paired fight on two levels, interactive and negative interaction:



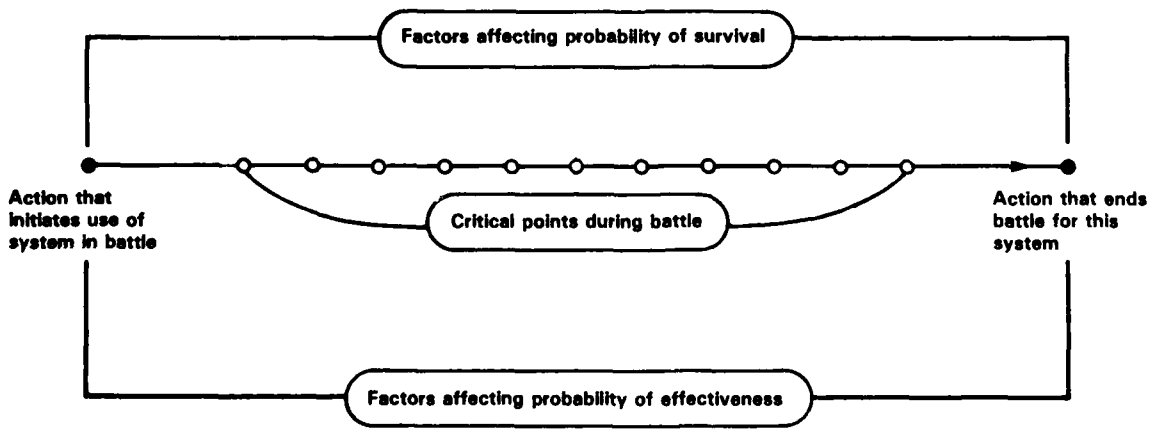
the interactions in a battle situation have to be understood before the test objectives can be clearly stated. In other words, understanding a specific battle situation in order to define the appropriate test objectives requires looking at military history, documentation, and data from exercises and at the records of developmental and operational testing and evaluation. The more complex that the battle situation is, the harder it may be to test it realistically. Moreover, there is always some constraint on realism in testing for combat operations and performance, making it impossible to attain some research objectives.

A complex battle situation can be understood by looking at its basic elements. Those shown in figure 10 are common to most battles--the levels of support, the pairings of forces, and the concepts of interaction. A well-defined and clearly stated test objective takes into consideration how the elements of the battle are likely to be related, as in the models in figure 10.

The decision to use a weapon system in combat depends considerably on what is known about the probability of survival in the struggle against the enemy and the probability of effectively deterring or defeating the enemy, as we show in figure 11. Test objectives should be defined to include the critical operational factors that affect these probabilities. Therefore, each weapon system that is included in a test should be examined in relation to the beginning and the ending and each critical point of its use during battle.

As a data source, military history is especially useful for JT&E, because it provides reminders that surprise and confusion are important variables in combat, not to be ignored in defining test objectives. When more than one service is to operate in one environment, reviewing the regulations and training manuals on doctrine, tactics, and operations that are issued by the JCS and

Figure 11
The Battle as the Interaction of Factors in the Employment of a Weapon System



the services helps determine the objectives that can be sought realistically. Intelligence documentation helps in deciding which threats to simulate. Data from military exercises generally reveal what has been learned about the critical operational features of weapons in combat. They are important in defining objectives, especially when tactics, doctrines, or procedures are in doubt. All these sources of information are particularly useful in understanding what critical issues cannot be tested.

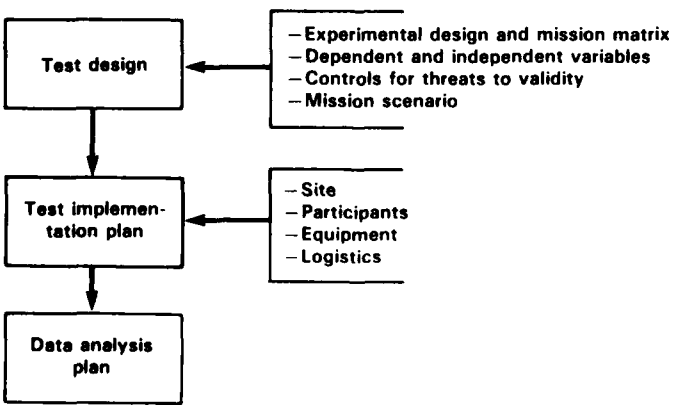
Deciding whether and how to construct simulations of battle situations for testing weapon systems and equipment also depends on the developmental research. A weapon system's potential limitations on the "battlefield" can be understood from the technical features that have been made evident in its developmental testing. Similarly, the operational testing and evaluation that have been done in the past reveal critical operational features and how they might be addressed. They may also reveal whether several tests have shared those critical factors in common and whether the data are compatible and can be automated.

In some instances, however, it may be impossible to determine the feasibility of addressing certain test issues. When this happens, the analyst must see whether a feasibility study has been planned, conducted, and evaluated before approaching the next step in the test process--planning the test. That is, going forward to coordinate, instrument, and measure the performance of many test participants should be based on a determination of exactly what is attainable within carefully identified boundaries of investigation. Step 2 of the test process, defining the objectives, entails giving specificity to the issues that are to be resolved by defining the scope of activities to be addressed.

Step 3: Planning the test

Step 3 involves a series of decisions about what to include in the test that have major implications for every participant.

Figure 12
Step 3: Planning the Test



As figure 12 shows, the decisions may take the form of a test design, a test implementation plan, and a data analysis plan.

The test design

For most JT&E's, a design or "mission matrix" spells out the kind and amount of data that will be required if the test's objectives are to be met. The design or matrix specifies the dependent and independent variables that are of greatest interest and the number of observations that have to be made of them. It also provides an indication of whether the data from the several test groups will be comparable, whether it will be possible to generalize from the test to combat situations, and how complex the test will be and how much it is likely to cost. When cost must be balanced against statistical rigor, it is useful to consider alternative designs.

The dependent variables in military testing often serve as measures of "survivability" or as measures of effectiveness in combat. The independent variables are what will affect combat performance. For example, if weather is being treated as an independent variable, "poor weather" and "good weather" will be explicitly defined in the test design in terms such as visibility, cloud cover, temperature, and wind speed. Accurate interpretation of the test results depends on the careful delineation of the variables.

In this report, we have called the potential sources of error in estimating combat performance "threats to validity" in testing--defining as "valid" that which measures what it was intended to measure. For example, the failure to consider the effect of the passage of time may threaten a test's validity. Pilots who fly trials in the evening may be hungrier and more tired than in morning trials they flew the same day; therefore, in analyzing step 3 of the test process for a JT&E whose objective is to compare weapon systems and tactics, one would look to see whether equal numbers of trials are scheduled for the morning and for later in the day for each set of conditions being tested. If this is not part of the design, differences in performance could be attributable to the condition of the test participants rather than to the factors specifically being tested. In some operational tests, however, the stress and fatigue of actual battle may be the factors to test. When they are, an analysis of the design would look to see whether all test comparison groups have equal representation of the conditions of stress and fatigue.

Another source of error or threat to validity in testing is the failure to consider the loss of test participants from comparison groups because they have been reassigned to other duty before the test is completed or suffer an accident during the test that prevents them from going on in the same way. Such losses can be controlled for in the analysis, if they are reported. Another threat is the effect missions early in the test can have on later ones. For example, in a test designed to examine each pilot's

first pass over enemy forces in battle, each pilot in the test should make only one pass, so that every pass is truly a first, unaffected by learning. Similarly, in a test of two weapon systems, the instruments that are used for recording the test data and the way they are used should be comparable. Otherwise, differences that are observed in the weapons may actually stem from the instrumentation.

Among other possible threats is a comparison of combat performance that is based on differences in selecting the test's participants, although such differences can be controlled for in the design or the analysis. For example, in a test comparing the effectiveness of an Army helicopter with an Air Force fixed-wing aircraft, one would look to see whether the pilots from the Army and the Air Force had been selected with the same criteria. If the Army sent its average pilots and the Air Force sent its best and brightest, observations about their "survivability" may be confounded by the differences.

The test implementation plan

A JT&E's test implementation plan defines how the test's joint missions will simulate battle. It proposes combined operating procedures for simulating the actual battle procedures to be expected from the military history and other documents reviewed at step 2 of the test process. For each level of battle being simulated, a specific "scenario" is written. Scenarios for ground, air, and sea forces account for the proposed enemy's equipment, tactics, and procedures, given current U.S., NATO, and other intelligence information. Where the test's scenarios depart from real combat situations, as when it is necessary to prevent the battle area from affecting nearby civilians, each difference is carefully documented.

Other elements of the implementation plan include documentation of the ways in which the test site does and does not represent the environment being simulated. Safety needs, ceiling limits, environmental conditions, and the availability of instrumentation systems and facilities may all cut down on the number of test sites that can be considered, but a test that is done in weather like Germany's, for example, may not indicate what military performance would be like in the Middle East. Details of the personnel are included--the numbers needed, the abilities they should represent, who is to have control over them, the flexibilities in their schedules, and so on. If the test is intended to assess the performance of both "friendly" and "enemy" forces, the plan may show how it is being arranged that the participants who play them never meet except in combat.

Equipment for both the weapon systems and the instrumentation systems is scheduled for use and noted in the implementation plan. When the equipment consists of developmental models, details of how they differ from final production models are included. The equipment that will be used to simulate enemy equipment must be

understood, through a delineation of the expected differences. The plan includes details of how the instrumentation systems are to collect the data, whether individuals will simply observe and record specific actions or elaborate electronic time-in-space information will be automated, and what their strengths, weaknesses, and possible effects on test results are.

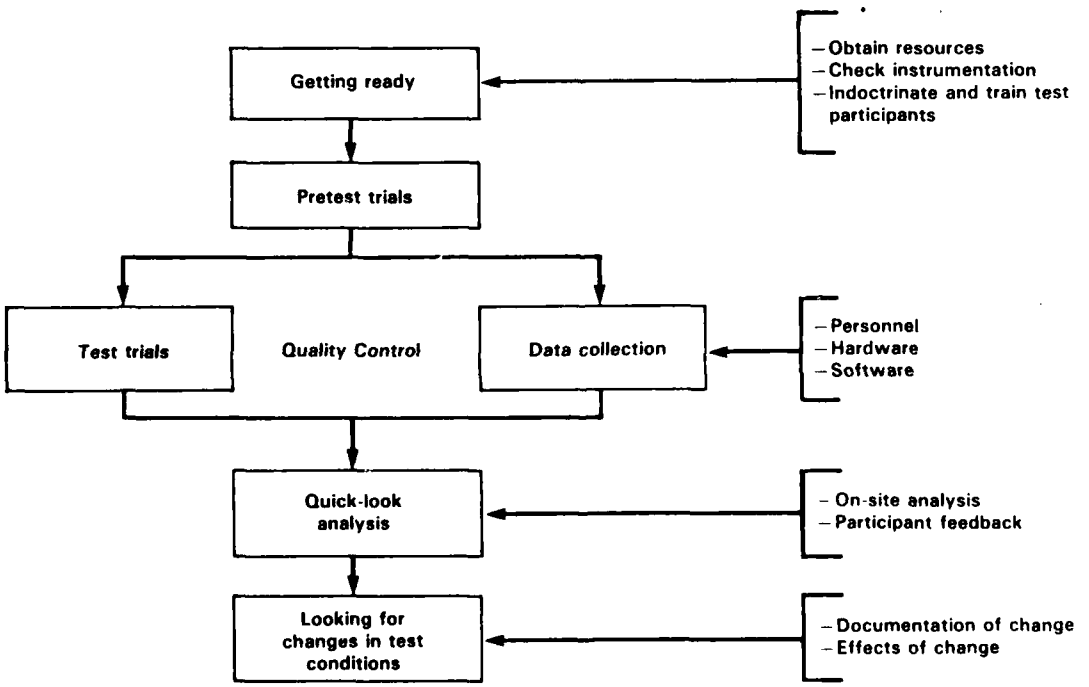
The data analysis plan

The data analysis plan is a formulation of how the test's data will be analyzed and evaluated. Its purpose is to specify how each of the test's objectives will be addressed in the analysis and how the analysis itself will be evaluated. It is also used in checking to see that all the necessary data are collected and that the estimates of time and effort required for analysis and evaluation are accurate. In other words, the data analysis plan establishes the specific criteria for judging the test's results and for deriving estimates of combat performance from it.

Step 4: Implementing the test

Implementing the test gets it under way, runs the trials, collects the data, and looks for threats to the results stemming from changes in the design, plans, or conditions of the test. Step 4 is outlined in figure 13. "Getting ready" means amassing the resources, checking the instruments and weapon systems, and

Figure 13
Step 4: Implementing the Test



training the test's managers and support personnel according to the test plan. More training can allow participants to become unduly familiar with the test area and may affect the test data. Checking the equipment helps anticipate deviations from error rates specified in the designs and plans and gives some indication of what problems will appear during the test's trials.

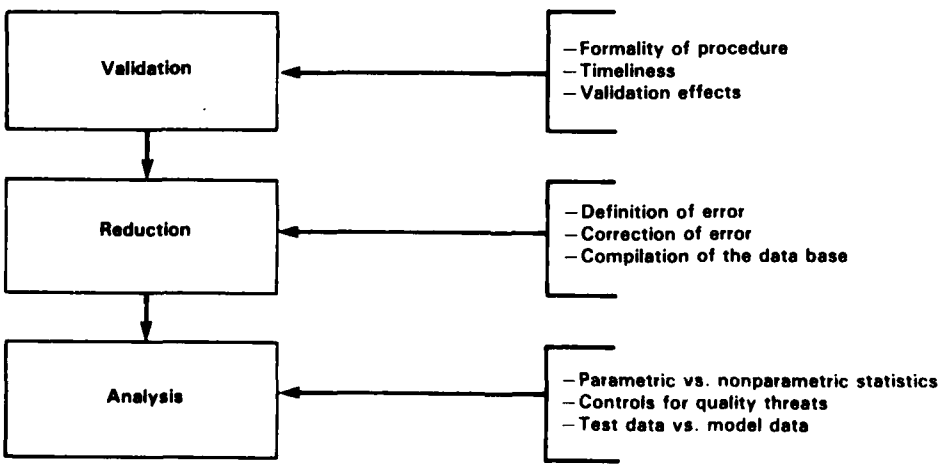
During implementation, the entire test procedure itself is tested, if possible, to identify potential problems before "for-the-record" testing begins. Pretesting trials permit a final judgment on the feasibility of completing the test according to design. Crises and equipment failures during pretesting sometimes lead to a revision in the test design or plans and a documentation of the changes. As the trials of the test proper are run, they are routinely monitored for equipment failures and corrections.

Step 4 also includes checking error rates systematically, holding debriefing sessions to help verify the data that are reported, and spot-checking the data collectors. Information from the test's participants about individual deviations from test rules or procedures helps the test's analysts interpret trends in the data, especially when they can monitor the test as it takes place.

Step 5: Analyzing the data

Step 5 depends on steps 1-4. Analyzing the data depends on knowing the test's context, objectives, plans, and implementation. Figure 14 shows the three parts of step 5: validation, reduction, and analysis. In some JT&E's, a formal "validation committee" screens all the test data--the passes over the battle area, the completion of missions, and the like--to determine that essential instrumentation and weapon systems were functional, that the test's procedures were followed, that the data are

Figure 14
Step 5: Analyzing the Data



sufficient and accurate, and that on these grounds each trial can be declared either "valid" or "invalid." The decision can be based on the judgment of the committee members or on standardized and systematic rules. In other JT&E's, this process of data validation may be less formal.

In either case, having established the rules for which trials will be counted before the trials begin, and following them rigorously, helps insure consistency in the data base. Trials that are declared invalid because they are faulty or incomplete are often omitted from further consideration, but when the validation criteria are not clear, then the invalid and questionable trials should be compared with valid trials in a search for significant differences. Notice that data that are declared "valid," or "invalid," in the validation procedure are data whose adequacy has been authorized and that the use of terms differs from ours in expressions such as "threats to validity," in which we refer to "valid" data as those that measure what they purport to measure.

In data reduction, the data are checked systematically for errors and omissions. If the data collection was appropriately monitored and documented, it is easy to find the problems in it. The rates at which data are missing should be compared across the variables of interest. In some cases, the test information may be reconstructed to account for partially missing data, but the reconstruction should be appropriately documented and the reconstructed data should be analyzed separately.

The analysis proper begins when the data base is complete. It follows the analysis plan that was written at step 3, searching for justifications for any deviations from it. Analysis beyond the plan might include controlling for threats to the quality of test results, as we explained at step 3. For example, if a pilot is really killed during the test, the trials earlier than the fatal event should be analyzed separately from the trials afterward, before they are combined for overall analysis. Statistical techniques appropriate for the analysis of operational test data should be used, and it should be made clear whether a balance has been struck between using the data actually collected during the test and relying on computer models (to estimate "probability of kill," for example). Analysis that depends on a model may be misleading if the model is an insufficiently realistic representation of combat or if the test data do not meet its assumptions.

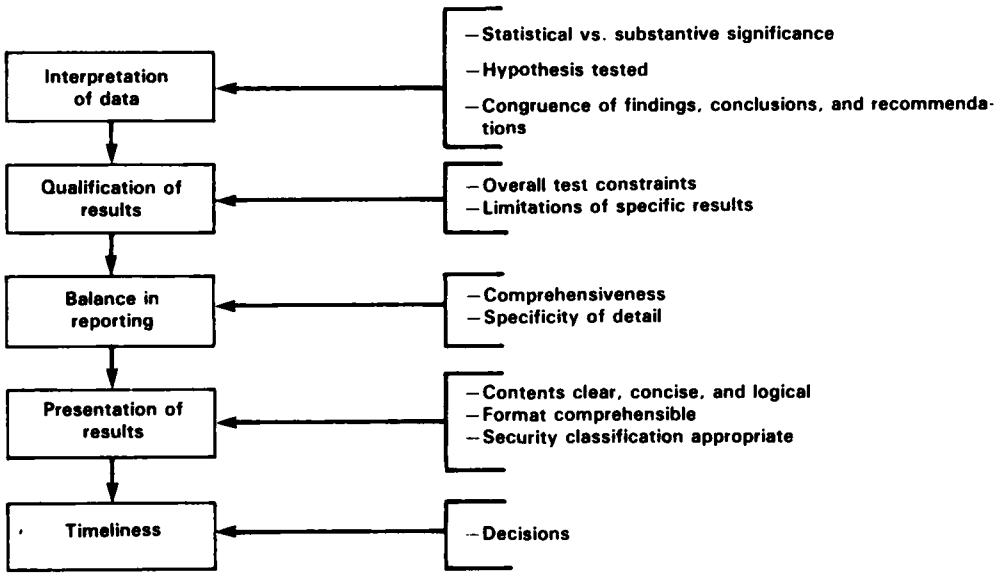
Step 6: Reporting the results

A test's report should reveal how hypotheses, c eria, and standards for analyzing the data, as set forth in the test design and plans, formed the basis for interpreting the data. Adequate interpretation both recognizes that statistical significance does not imply substantive significance and follows the logic by which test data lead to conclusions and conclusions lead to recommendations. The report should explain how the testing situation was constrained--by the infeasibility of testing certain issues,

limits in the instrumentation and equipment, crises during implementation, assumptions required during analysis, and so on. For example, if an accident during a test trial led aircrews to change their behavior significantly, the ways in which this affected the analysis should be discussed.

The report should be comprehensive and adequately detailed. Technical appendixes should be reserved for supplementary, not primary, information, and the entire presentation should be clear, concise, logical, and organized in a way that meets the requestor's needs. It should be timely and classified at the appropriate security level. An open version of a secret report can preserve national security while making the findings more widely available. Figure 15 summarizes step 6.

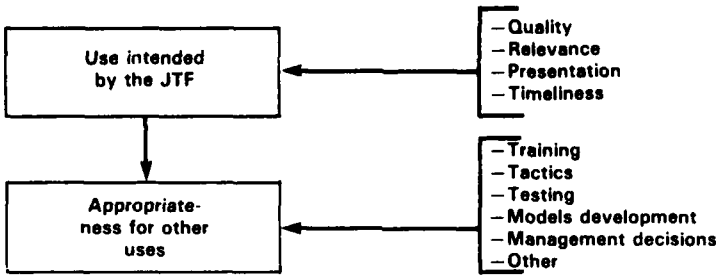
Figure 15
Step 6: Reporting the Results



Step 7: Using the results

In our analysis of JT&E, we looked at the way in which the reported results had been used as an index of their usefulness. We thought it important at step 7 to understand the intentions of the users who requested the tests and also to know what unintended utility the tests had, as we show in figure 16. The quality, relevance, timeliness, and presentation of results, the product of step 6 in the test process, help determine whether a JT&E is useful to those who requested it. However, the proposed use of the results of a test is understood as part of its context, as we saw at step 1, so that how its original objectives were modified at steps 2-5 is also an indicator of the test's usefulness.

Figure 16
Step 7: Using the Results



That is, intentions may change, depending on whether the accountability, knowledge, or management perspectives of a test's requestor have been addressed. While a test's results may be intended for developing training, tactics, weapon systems, models, other tests, and the like, uses beyond those that the requestor proposed may be anticipated in the test's final report, in supplementary reports by the participants on what they learned, in service memorandums, and in reports of subsequent efforts to use the test data. Evidence of how the test results are later used, and whether they are useful, may be found in the regulations and training manuals of the armed services, among other places. How appropriate any given use is should be judged against the quality of the reports and the degree to which the data may have been distorted or misinterpreted.

ASSESSING TEST QUALITY

In assessing the quality of JT&E, we used the phases shown in figure 6. Having used the seven steps of the test process to identify specific threats to the quality of IIR Maverick, TASVAL, and ACEVAL, then we tried to determine how the quality of these tests might have been lowered. For example, in testing the ability of the Army and Air Force to provide close air support jointly, not playing the friendly ground force to oppose the enemy ground force would reduce pilots' workload, since they would not have to distinguish friend from foe. The aircraft could engage at greater standoff ranges, which could result in an overestimation of the ability of friendly aircraft to survive enemy air defenses. Similarly, in a test with no definitive criteria for excluding trials from the data base (because their data are inadequate), reported outcomes could vary from mission to mission with no way of pinpointing when or by how much.

That is, to assess each test's quality, we tried to determine the credibility of the estimates of combat performance that were derived from it. In this sense, finding high quality, or credibility, meant that we found that the test's results were both reliable, or that they could be repeated under similar circumstances, and valid, or that they did in fact measure what it had been claimed they measured.

We tied our reasoning to the situation of each case study (as we report in chapters 4, 5, and 6), but we also tried to trace the threats to quality back through the seven steps of the test process. Since it is often not possible to know beforehand how well specific combinations of weapon systems, force structures, and so on will perform in war, operational testing and evaluation aim to produce accurate estimates of combat performance from mock combat that is made as realistic as can be. Avoiding underestimates and overestimates and misleading reports of credible estimates depends on accounting for all the conditions, events, decisions, and changes that occur throughout the test process. Therefore, in making our assessment of quality, we examined the success with which the tests met their objectives without serious damage, from all possible threats, and we judged how well they allowed accurate projection, from the testing range to the battlefield, of what is likely to take place in combat.

ASSESSING THE USEFULNESS OF TEST RESULTS

Examining whether the results were useful and how they were used completes the phases of our analysis of cases, as shown in figure 6, as well as our conceptual framework for analyzing the test process, as shown in figure 7. That is, as we have explained throughout this chapter, and especially at step 7 of the test process, we based our judgments about the use and the usefulness of JT&E results on our review of quality as it is affected by the factors associated with a test's context, objectives, plans, implementation, analysis, and reports. We took high quality to be a prerequisite for usefulness, believing that test results of low quality might be too erroneous to be credible. This would also imply that almost any unintended use would be inappropriate.

We did not take quality as a guarantee that, if it were high, a test would be either useful or used. A test that is high in quality, by being methodologically sound and accurate, will probably be useless and unused if it is irrelevant, does not fulfill the requestor's need for information, bears little relation to or is reported too late for the decisions it is needed for, or is not presented with thoroughness, balance, and clarity.

SUMMARY

The method we have outlined in this chapter permitted us to make detailed assessments of a large number of test variables and conditions. Our conceptual framework allowed for insight into the strengths and weaknesses of the JT&E process, and it allowed for inferences about similar test processes. Its limitation is that we cannot generalize from it to all test-and-evaluation approaches. Because we reviewed only three tests, our findings are not necessarily representative of all tests, and it cannot be assumed that they apply to all JT&E's.

Nevertheless, we believe that our findings about test quality and usefulness may be more valuable to test users and managers

than a survey of all joint testing and evaluation, because we have been able to describe the relationship of each step of the test process to subsequent steps. In our review, we considered longitudinal information, so that each test can be described in its entirety in terms of its context, its development, and its implementation. In addition, our seven-step framework for the test process made it possible to gather the perspectives and assess the knowledge of the various groups that were involved with the three tests we examined. We attempted to consider all aspects of each test in order to diminish the likelihood of bias and broaden our data base. We tried to make a full description of the three test situations, despite the diversity of the characteristics of JT&E. We attempted to obtain sufficient information to explain what is common in the quality and usefulness of all three case study tests--IIR Maverick, TASVAL, and ACEVAL--and what is unique in each one.

CHAPTER 4

THE IMAGING INFRARED (IIR) MAVERICK JT&E

In February 1977, a joint operational test and evaluation of the Imaging Infrared Maverick, a heat-seeking, air-to-surface missile, was conducted at Ft. Polk, Louisiana. U.S. Air Force pilots flew single-seat aircraft and simulated the launching of the missile against enemy ground forces. The purpose of the test was to determine the operational capabilities and limitations of the IIR Maverick system. It was hoped that the findings would clear up operational uncertainties that had been identified by OSD officials who had decided that the missile was ready for full-scale engineering development but who were not convinced of its operational feasibility. It should be noted that this joint test did not serve the primary purpose of the JT&E program, in that it did not examine the IIR Maverick weapon system's performance in a joint environment. Instead, it focused on the secondary purpose of the program, in that it examined the operational requirements of the IIR Maverick weapon system.

As a result of the test, the joint test force (JTF) concluded the following:

> "All major goals of the IIR Maverick JOT&E [joint operational test and evaluation] were achieved. The JOT&E answered the critical concerns which resulted from the DSARC [Defense Systems Acquisition Review Council] II deliberations. The operational test data, with the modifications planned in going from Advanced Development to Engineering Development design, indicate that the IIR Maverick should meet operational requirements and thus support the transition of the system to Engineering Development.
>
> "Overall, the JOT&E demonstrated impressive capabilities for the IIR Maverick in a highly realistic environment. The IIR Maverick gives the Maverick family autonomous night attack capabilities (once the system has been cued to the target area) and improved adverse weather performance, capabilities needed by the Tactical Air Forces to counter the massive armor threat of the Warsaw Pact." (II.C.21, p. iv, emphasis added)[1]

In this chapter, we evaluate these and other conclusions put forth by the JTF by examining the test results according to the approach we discussed in chapter 3. We present evidence that demonstrates that all the major goals of the JT&E were not, in fact, achieved. Our examination of the operational test data shows that

---

[1]The bibliographic data for all quotations in this chapter are in appendix II, section C, which contains our references to documents on the IIR Maverick.

they do not indicate that the IIR Maverick missile met its operational requirements. We also question whether the test environment was indeed highly realistic. We find that the test did not compare the imaging infrared version of the missile with other missiles of its type to provide evidence of the IIR Maverick's improved capability. Although we question the conclusions put forth by the JTF, we find that other aspects of this joint test were done well and provide valuable lessons for future tests.

We focused on the information in the joint test force report. We also include applicable information from the analyses conducted and reported by the System Planning Corporation, the U.S. Air Force Studies and Analyses group, and the Joint Services Electro-Optical Guided Weapons Countermeasures Test Program.

In the five sections in this chapter, we first provide information about the first step of the test process: the context in which the JT&E took place. Second, we describe briefly the test objectives and design. In the third and most lengthy section, we present our observations about the major threats to the test quality for all seven JTF objectives (see figure 17). For each objective, we reiterate the JTF's original statement of it and the conclusions as they were originally stated, elaborate on how the JTF addressed the objective and reported the results, and discuss the

Figure 17

The IIR Maverick Test Objectives

Objective	To provide data on	Pages
Transition	the operational difficulties associated with the transition from the navigational phase to the point in the attack at which the IIR Maverick is launched by day and by night and under limited visibility.	36-46
Valid target	the ability of the operator to interpret a valid target in the presence of battlefield clutter.	46-53
Cueing	the requirement for cueing.	53-57
Survivability	the extent of exposure to forward-area air defenses while accomplishing the functions that are necessary in delivering the IIR Maverick missile.	57-61
Single-seat aircraft	the ability to accomplish the IIR Maverick delivery function in a single-seat aircraft under operational conditions.	61-64
Countermeasures	the system's utility in the presence of countermeasures.	64-67
Thermal character	the thermal characteristics of the proposed targets.	67-71
	the thermal characteristics of the battlefield.	

major problems at each step of the test process--omitted issues (step 2), unrealistic test conditions (step 3), test changes (step 4), analysis problems (step 5), and reporting problems (step 6). Then we summarize our observations about test quality and conclude the chapter with a section on the final step in the test process, in which we make some observations about the usefulness of the test results.

THE CONTEXT OF THE IIR MAVERICK TEST

On September 28, 1976, the Defense Systems Acquisition Review Council II reviewed in detail the Advanced Development Imaging Infrared Maverick Program--a program for developing an imaging infrared "seeker" for the Air Force's Maverick missile, a precision air-to-ground weapon for attacking targets such as tanks. (See appendix IV: item 1 is a chronology of the JT&E and item 2 is a description of the missile.) The DSARC II assessed the readiness of the IIR Maverick for full-scale engineering development.

One of the primary purposes of a DSARC II review is to insure that the uncertainties in a system have been identified and that the risks that stem from them are acceptable. In this case, many operational uncertainties about the IIR Maverick were identified, but it was unclear whether their associated risks were acceptable. For example, the Director for Defense Test and Evaluation noted that further testing was needed while also recommending that the system be moved to the next phase of development, despite the uncertainties that had been identified (app. IV, item 3).

The DDT&E listed the IIR Maverick's operational uncertainties (app. IV, item 4) and acknowledged three testing options: (1) a test run by the Air Force under the guidance of the IIR Maverick program's manager, (2) a test run by the Air Force's independent Test and Evaluation Center (AFTEC), and (3) a "mini"-joint operational test to be conducted as a quick response. (He also noted that the Joint Services Electro-Optical Guided Weapons Countermeasures Test Program would be asked to support the planning and execution of the test and to report on the susceptibility of the system to countermeasures, regardless of who conducted the test.) The DDT&E appeared to favor the third approach because "the design and analysis of test results could be done by an independent contractor who has no self-serving interest" and "the independent analysis would lend more credence to the test findings" (II.C.5, p. 2).

On October 14, 1976, the Assistant Director for Tactical Systems Test and Evaluation stated in a memorandum to the DDT&E that the planned Air Force tests of the IIR Maverick would not resolve operational uncertainties in certain areas (app. IV, item 5). Because of these concerns, the joint test approach was selected.

On November 19, 1976, the Deputy Secretary of Defense issued the DSARC II decision memorandum on the IIR Maverick, stating that the DSARC had

"found that a very extensive test program has been conducted and that the basic technical feasibility has been demonstrated. The DSARC has expressed the need for further operational testing to understand more fully any operational uncertainties or limitations which may exist and to facilitate the evaluation of appropriate operational tactics during the next phase of the program." (II.C.1, p. 1)

The program's transition to full-scale engineering development was approved conditionally. An operational test would have to be conducted with measurements of thermal clutter (that is, various sources of heat other than targets) under battlefield conditions as realistic as practicable and including countermeasures. There would have to be a DSARC review of the program and its testing progress before the pilot production of 240 missiles. The Deputy Secretary of Defense further specified that partial test results would have to be made available by mid-March 1977 and that the final report would have to be available to OSD by August 1, 1977.

The joint test program was initiated immediately. On November 26, 1976, the Director for Defense Research and Engineering issued a memorandum to the secretaries of the military departments in which he established the IIR Maverick JT&E. The Air Force would be the lead service and work jointly with the Army and the Navy. The memorandum also provided milestones, confirmed DDT&E funding for costs unique to the test, and named the System Planning Corporation (on retainer to DDT&E) to assist in planning, monitoring, and reviewing the test and to conduct an independent evaluation of it.

THE TEST OBJECTIVES AND DESIGN

The purpose of the IIR Maverick test was to provide data so that the operational uncertainties of the IIR Maverick that had been identified during the DSARC II deliberations could be more fully understood. (See appendix IV, item 6, for a more detailed description, and item 7, for a list of the uncertainties.) The test had the seven specific objectives that we presented in figure 17. The thermal character objective was further divided into two objectives, as the figure shows. The design matrix that was proposed for addressing these objectives is shown in figure 18 (on the next page). The design called for 24 missions (each cell in the matrix is a mission), with at least 6 passes during each mission. Eight dependent variables were proposed as indicators of the system's performance in the test, and they are listed in figure 19 (on the next page) along with the independent variables. A summary of the JTF's original data analysis plan is in appendix IV (item 8).

In the test, which was to simulate the weather and battlefield conditions of combat in a midintensity conflict in central Europe in 1982, the IIR Maverick missile was to be used as a standoff air-to-surface weapon against enemy armor and air defense

Figure 18
Design Matrix for the IIR Maverick JT&E

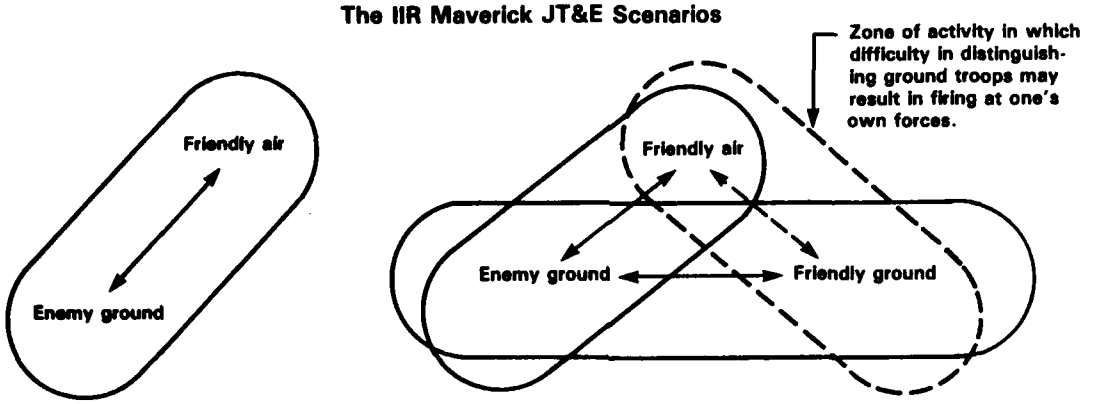
Aircraft	A-7				A-10				F-4E			
Test scenario[a]	CAS		PPI		CAS				PPI		PPS[b]	
Acquisition aids[c]	INS + FAC		INS		FAC		FAC + Pave Penny		Pave Tack		Pave Tack	
Visibility[d]	Good	Poor	Good	Poor	Good	Poor	Good	Poor	Good	Poor	Good	Poor
Day-midday	–	1	1	–	1	1	1	1	–	1	1	–
Night-dusk	–	1	1	–	1	1	1	1	1	–	–	1
Night-midnight	1	–	–	–	1	–	1	–	1	–	–	–
Night-predawn	–	–	1	–	1	–	1	–	–	–	1	–
SUBTOTAL	1	2	3	–	4	2	4	2	2	1	2	1
TOTAL	3		3		6		6		3		3	

[a]CAS = close air support; PPI = preplanned interdiction; PPS = preplanned strike.
[b]These missions were dropped from the test.
[c]INS = inertial navigation system; FAC = forward air controller; Pave Penny = a sensor for acquiring laser-designated targets; Pave Tack = a forward-looking infrared system.
[d]Poor visibility = less than 5 statute miles; good visibility = 5 statute miles or more.

Figure 19
The Major Variables Considered in the IIR Maverick JT&E

Independent variable	Dependent variable
Type of strike aircraft A-7 A-10 F-4E (dropped from the test) Acquisition aid (cues to the target area and targets) Inertial navigation system (A-7 only) Forward air controller Pave Tack (F-4E only) (dropped from the test) Pave Penny (A-10 only) Target scenario Close air support Preplanned interdiction Preplanned interdiction strike (dropped from the test) Visibility Poor (less than 5 statute miles) Good (5 statute miles or more) Time of day Midday (10:00 a.m. to 5:00 p.m.) Dusk (between sunset and one hour past sunset) Midnight (10:00 p.m. to 2:00 a.m.) Predawn (between one hour before sunrise and sunrise)	Probability of target-area acquisition Range of target-area acquisition Target detection range Target lock-on range Launch and abort range Probability of attacking a valid target Time from wings level to launch and abort Probability of aircraft survival

Figure 20
The IIR Maverick JT&E Scenarios



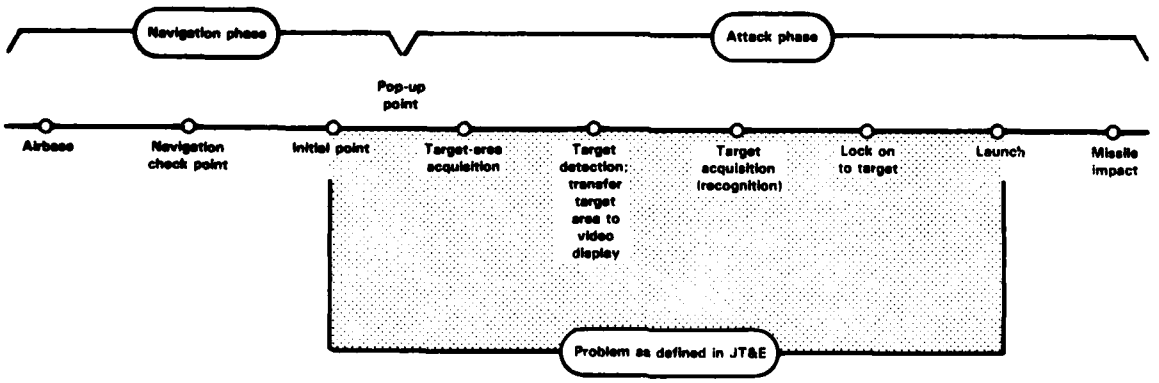
Preplanned interdiction: one-paired interactive fighting on two levels

Close air support: two-paired interactive fighting on two levels with negative interaction

units behind the forward edge of the battle area. There were to be two scenarios--close air support and preplanned interdiction --which we illustrate schematically in figure 20.

The process of employing the IIR Maverick is shown in figure 21. In the navigation phase, a pilot flies from an airbase to pop-up point; in the attack phase, the pilot acquires a target area, transfers the target area to the infrared video display, detects and acquires a target, locks onto the target (that is, automatically puts the target in the missile's field of view), and launches the missile. The pop-up point is the point of transition from navigation to attack. The points in the process of employing the IIR Maverick before reaching the initial point during navigation and after launching the missile during attack were not considered in the JT&E.

Figure 21
The Process of Employing the IIR Maverick



In close air support in combat, the IIR Maverick is to be employed in two-paired interactive fighting between air and land forces. That is, friendly air forces and ground forces will both be shooting at enemy ground forces. Negative interactive fighting is also possible--that is, friendly air forces might shoot at friendly ground forces where friendly ground troops are close to the enemy's troops. Nevertheless, the possibility of shooting at one's own troops was not considered in this test. In interdiction in combat, the IIR Maverick is to be used in one-paired interactive fighting on two levels--that is, friendly air forces will shoot at enemy ground forces and vice versa.

Since battle interaction is difficult to simulate in testing, it is often addressed through analysis. In this test, the probability of aircraft survival was estimated with computer models. They used test-generated data and theoretical data for the performance of the aircraft employing the IIR Maverick missile and theoretical data for the expected performance of the enemy air defense system.

During the test, 23 "for-the-record" A-10 and A-7 missions were flown, and there were 105 record passes. Thirteen additional record passes were for the purpose of examining the countermeasures objectives. According to the joint test force, 58 percent of the record passes, or 61 passes, were flown during the day or at twilight and 42 percent, or 44, were flown at night (defined as "one half hour after sunset to one half hour before dawn"). Of the 105 record passes, 14 percent, or 15 passes, were flown with ground visibility of 1 to 3 miles; 37 percent, or 39, were flown with ground visibility of 4 to 6 miles; and 49 percent, or 51, were flown with ground visibility of 7 to 9 miles.

THE QUALITY OF THE TEST RESULTS

In the seven sections under this heading, we examine each of the test objectives listed in figure 17 in terms of how the omission of issues, unrealistic test conditions, test changes, and problems in analysis or reporting affected the quality of the test results. All the quotations of the JTF's objectives and conclusions that we display at the opening of each section are from the official report of the IIR Maverick JT&E issued by the U.S. Air Force Test and Evaluation Center. (The objectives are all on page II-1 and the conclusions are all on pages II-7 through II-9 of the JTF's report, unless noted otherwise; see document 21, section C, in appendix II of our report.)

Elaboration of test objective and reported results

The transition objective addressed the operational difficulties associated with using the IIR Maverick during the day, at night, and when visibility is limited. The transition from navigation to attack (as depicted in figure 21) is the point at which the pilot, having flown past the initial point, "pops up" by

TRANSITION OBJECTIVE

JTF objective

Evaluate the IIR Maverick with respect to the "Operational difficulties associated with transition from the navigational phase to the point in the attack phase when the IIR Maverick is launched under day, night and limited visibility conditions."

JTF conclusions

"Many of the DSARC reservations regarding the combat utility of the IIR Maverick focused on the operational requirements associated with transitioning from the navigational phase of the mission to the point in the attack phase when the missile is launched. Using current tactics, procedures, and onboard systems, the JOT&E demonstrated that transition is not a problem for conditions similar to those tested. The pilots used realistic [forward air controller] information and onboard navigation systems to navigate accurately from the [initial point] to the pop-up point, the point of transition from navigation to the attack. Steering information from the A-7 [inertial navigational system], A-10 Pave Penny, and visual cues from the realistic battlefield proved to be sufficient aids for placing the targets within the IIR Maverick field-of-view. Since the A-10 is not currently [inertial navigational system]-equipped, it is important to select prominent [initial points], particularly at night, which can be easily located by [dead-reckoning] navigation and/or onboard systems such as TACAN."

bringing the aircraft to a higher attack altitude at a given time and distance. Thus, the operational difficulties that were posed in the test were associated with the pilot's ability to find a target area from the attack altitude and then successfully acquire and lock onto the enemy target and launch the missile.

The JTF concluded that making the transition from navigation to attack was not a problem in the JT&E given current tactics, procedures, and systems aboard the aircraft. The JTF reported that

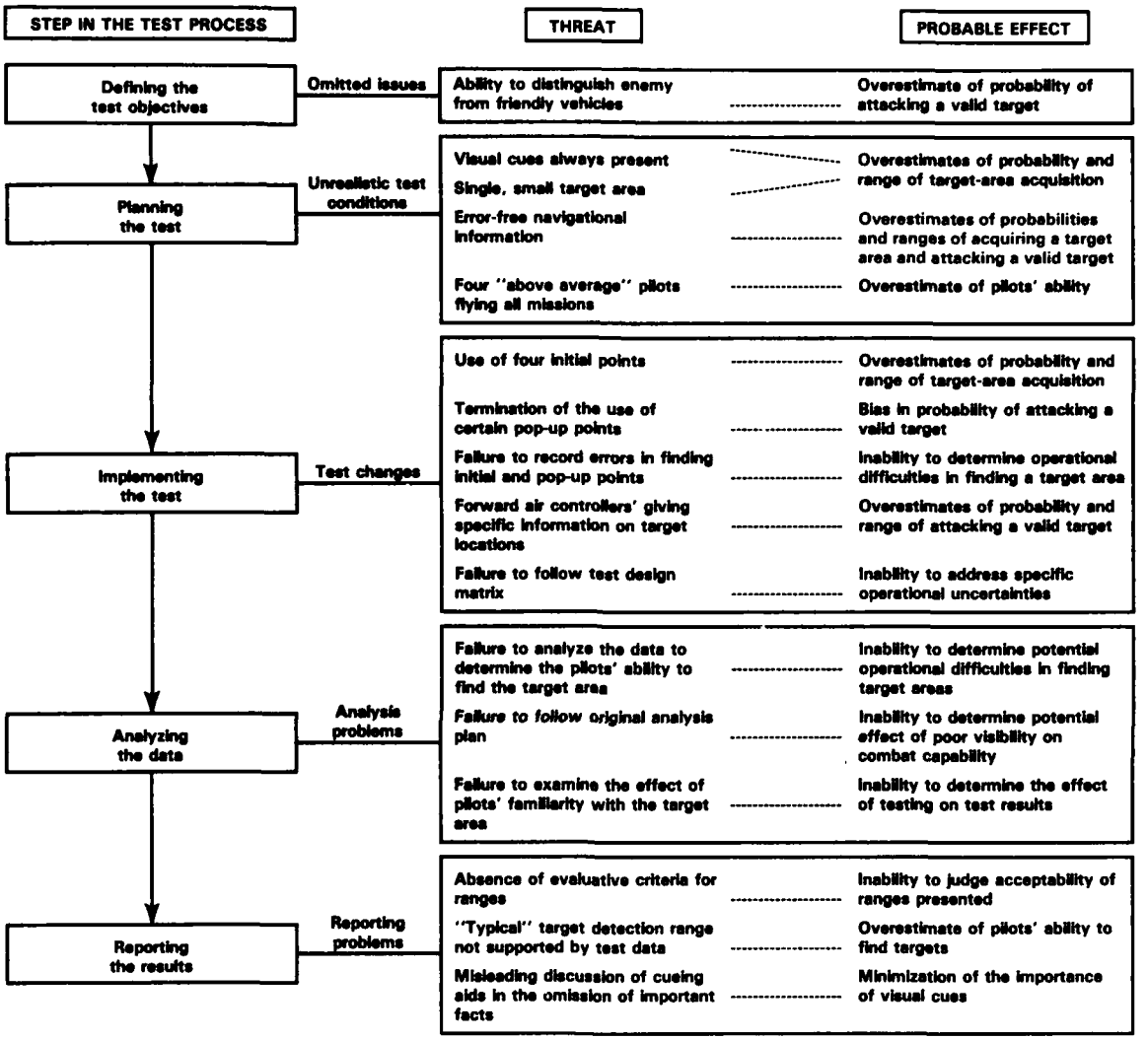
> "of the 105 usable record passes, the pilots did confine their search for specific targets with the cockpit display to the immediate target area on all but two occasions." (II.C.21, p. II-13)

and that

> "The median value of target area acquisition range was observed to be with 10th and 90th percentile points at Depending on ingress altitude and visibility conditions, pilots were sometimes able to acquire the target area visually before reaching the pop-up point." (II.C.21, p. II-19)

The A-7 pilots reported that they had little difficulty during navigation because their navigation system, the "inertial navigation system," provided the necessary guidance. However, the A-10 pilots, lacking a navigation system, did report difficulty in finding less prominent initial points, especially at night and when visibility was poor.

Figure 22
Threats to Test Quality: The Transition Objective



Threats to test quality

In figure 22, we show the major threats to the quality of the test results relative to the transition objective in the same sequence in which they occurred, chronologically, during the JT&E. There were threats at all five steps of the test process from defining the test objectives through reporting the results.

Omitted issues

Although it is recognized that all issues cannot be tested, the failure to at least discuss the critical issues that were not tested is a problem. The problem of distinguishing one's own

ground forces from those of the enemy was omitted in defining the objectives of this test. No friendly ground forces were simulated in the test. The Air Force Manual defines close air support as "air action against hostile targets in close proximity to friendly forces" and adds that it "requires detailed integration of each air mission with the fire and movement of those forces" (II.C.6, p. 6-1). This suggests that differentiating between friendly and enemy vehicles could be difficult in close air support missions and may result in operational difficulties, but omitting the issue resulted in no information about how the problem might affect the ability to use the IIR Maverick. In addition, the JT&E results probably overestimate the probability of attacking valid targets because the pilots did not have to distinguish enemy forces from friendly forces.

Unrealistic test conditions

Four test conditions that were established at the planning step make it difficult to generalize from the test results on transition to performance in combat. The conditions were unrealistic in that the pilots had (1) visual cues, (2) a small target area, (3) error-free navigational information, and (4) "above average" skills.

First, it was recognized in the test concept that visual cueing aids may not always be available in combat, particularly since the IIR Maverick is meant to be employed at night and in poor weather; nevertheless, the test was designed so that visual cues were always available. The presence of visual cues was a criterion for calling a pass "for the record." In the close air support scenario, the tanks had to be firing; in the preplanned interdiction scenario, six hulks located at six predetermined spots had to be lit and burning. The ability to acquire a target area without these unique visual cues was not addressed. Given the proposed use of the IIR Maverick in poor weather, the JT&E results probably overestimate the ability of pilots to find target areas and the range at which they can find target areas.

Second, the use of a single, small target area may have precluded the emergence of some of the operational difficulties of employing the IIR Maverick. The System Planning Corporation reported that two areas on the testing range could have supported the target array, but the only one that was used measured 1.5 kilometers by 2.0 kilometers. Since the pilots acquired the same target area over and over again, they became quite familiar with it. Their unrealistic familiarity probably means that the test data overestimate the probability and the range of acquiring a target area.

Third, the information that was given to the test pilots to help them navigate to the initial point and find the target area was free of error. Thus, the test did not simulate three types of error that might be expected. (1) The pilots had a navigational aid called "TACAN" to help them locate initial points. According

to test documents, the Air Force has and probably will deploy in combat portable TACAN stations that the enemy probably will destroy quickly or suppress. (2) The A-7 in the test was equipped with an inertial navigation system that is known, because of the frequency and quality of its update information, to be prone to error. However, the test pilots' ability to update the system accurately for an attack was not tested. (3) In the preplanned interdiction scenario, a pathfinder aircraft was simulated by marking enemy forces visually with a burning vehicle; accurate coordinates for the enemy target were successfully and accurately communicated to the attack aircraft in real time. However, the test ignored the possibility of delay between the pathfinder's observing and relaying information about where the IIR Maverick pilots are to find the enemy target and the enemy's subsequent movement. In summary, data on the possibility of navigational error from these three sources were not collected, and the failure to simulate them in the test may mean that the combat capability of the pilots employing the IIR Maverick was overestimated.

Fourth, since the same two A-7 pilots and the same two A-10 pilots flew all pretest and all test missions, they may have learned how to overcome some operational difficulties as early as during the pretest missions. Moreover, although the JTF provided no specific information on how the experience of these pilots compared with that of other pilots likely to fly the IIR Maverick, it did note that the four test pilots were considered to be "above average." Since the IIR Maverick may be employed by pilots with less experience and skill than they had, the JT&E results may overestimate the pilots' ability to successfully use this missile.

Test changes

Not all the components of the test plan were implemented as designed. One that was not had to do with test changes related to the pilots' ability to find the target area and the targets. Another involved changes to the test mission matrix.

The pilots' ability to find the target area and targets was an operational uncertainty in the employment of the IIR Maverick. Four test changes may not have cleared it up. (1) Only four of six initial points were used. (2) The use of certain pop-up points was terminated. (3) Errors in finding initial points and pop-up points were not recorded, although it was planned to record them. (4) The forward air controllers provided specific information on the location of targets. These conditions in the implementation of the test had not been part of its design.

Six initial points--the points from which the aircraft approached the target area--had been chosen for the test, but only four were used. Three were to the west and one was to the east of the target area. No explanation for this change was provided in the test reports. The smaller number of initial points may have led to the pilots' gaining familiarity about them; this may have

led to an overestimation of their ability to find the target area. Consequently, the test may not have exposed whatever operational difficulties a pilot may have in finding target areas in combat, where the pilot has no choice about the location of the enemy forces and has to approach the target area from a less familiar initial point.

For bringing the aircraft up to a higher attack altitude after navigating from the initial point, the pilots were directed to any one of 33 different pop-up points during the initial stages of testing. During the test, the availability of some of them (the exact number was not reported by the JTF) was terminated because the pilots could not locate the targets with the missile from these altitudes. In combat, however, it may be necessary to begin an attack with the missile from some point at which the target cannot be located. This change in the test conditions may have biased the test results. In addition, the passes for which the terminated pop-up points had been used were not discarded or examined separately but were combined with all the others, so that no analysis was made of their effect or of the effect of terminating them.

The JTF reported that the A-7 pilots had little difficulty in finding designated initial points whereas the A-10 pilots, who did not have the inertial navigation system, had difficulty finding the less-prominent initial points, especially at night and when visibility was poor. Despite this finding, no quantitative comparative analysis was made. The JTF had planned to collect data on errors of latitude and longitude in finding exact initial points and pop-up points, but no such data were reported. Therefore, it is not possible to determine what specific operational difficulties there may be in finding initial points and pop-up points while attempting to use the IIR Maverick in combat.

The test plan specified that the information the forward air controllers were to give the pilots would be abbreviated in order to simulate a battlefield environment:

> "The FAC [forward air controller] will have realistic information available to him, that which he would normally be provided to request CAS [close air support] and could gather by observation from his ground position. He normally would not know, and should not brief the aircrew on exact maneuvers and precise tactics and activities of the enemy ground forces." (II.C.13, p. C-4)

In the test, the commencement of ground activity was based on predicted pop-up times, so that the forward air controllers had no information on exact ground maneuvers and activities until the predicted pop-up time. However, according to the JT&E documentation, they gave the pilots information such as the coordinates of forks in the road and which way the tanks moved along them. Moreover, according to the final test report,

"The FAC directed the strike aircraft to an IP [initial point] . . . , gave the pilot an ingress heading and time to a pop-up point, target direction and distance from the pop-up point, and a brief description of anticipated visual cues and target activity." (II.C.21, p. II-5)

It is not clear whether any of this information was realistic, but it was probably better than the test plan had specified. This change in the test conditions could have resulted in an overestimation of the pilots' ability to acquire valid targets.

The design matrix for the IIR Maverick JT&E that we showed in figure 18 was not completely followed. The test concept paper for this JT&E pointed out that

"because of the nighttime and adverse weather capabilities of the IIR Maverick the aircrew will be required to accomplish the attack sequence without the usual visual cueing aids available in daytime/fair weather." (II.C.5, att. p. 1)

To address this critical issue, the test design required that one third (or roughly 33 percent) of the missions be flown under conditions of poor visibility and two thirds (or roughly 66 percent) be flown under variations of night conditions (app. IV, item 9). Among the actual test missions, 22 percent were flown with poor visibility and 61 percent were night missions. Thus, a greater proportion of the actual test missions (which sometimes resulted in as many as 10 passes per mission) were flown with good visibility and during daytime than had been proposed in the test design. While the weather's effect on visibility cannot be controlled, the time of day at which missions are flown can be.

The test concept paper indicated that the "usual visual cueing aids" may not be available under certain conditions of IIR Maverick employment. The test design specified that six missions were to be flown in the F-4 (a two-seat aircraft), with the Pave Tack (a forward-looking infrared system) as the only cueing aid, but these missions were not conducted. According to the JTF final report, this phase of the test with the F-4 was deleted because the Director of Defense Research and Engineering decided that the ease with which the single-seat A-7 and A-10 pilots had employed the IIR Maverick warranted the deletion, and so did the substantial cost of moving the test site to another location in order to conduct the F-4 missions. However, according to the Air Force, using the IIR Maverick with the F-4 Pave Tack was one of the principal operational concepts for overcoming conditions of poor visibility when cues on the ground are not visible to the pilot. This means that one of the critical issues that had been identified for the IIR Maverick was not tested.

Analysis problems

The JTF analysis did not (1) examine the pilots' ability to find the target area under various conditions, (2) follow the

original analysis plan, and (3) determine what effect the pilots' becoming familiar with the target area had on the data.

The JTF report dealt only generally with the pilots' ability to find the target area, reporting that

> "On every record pass, the pilots believed that they had acquired the correct target area. Review of video tapes confirmed that, of the 105 usable record passes, the pilots did confine their search for specific targets with the cockpit display to the immediate target area on all but two occasions." (II.C.21, p. II-13)

The JTF acknowledged no specific problems in finding the target area. However, a more detailed analysis by the System Planning Corporation showed that the pilots could not acquire the target area in four instances and consequently aborted the mission.

Two of the four instances in which the pilots could not find the target area were during a preplanned interdiction mission. On this mission, the A-7 pilot had been successful in locating the target area on the first four passes, but after daybreak, when the burning hulks were less visible, the pilot was not able to pinpoint the target area, in spite of accurate navigation information. Since this was the only preplanned interdiction mission flown at sunrise, it represents all test passes flown right after daybreak, when visual cues may be little apparent.

The JTF did not examine one other indicator of the pilots' ability to acquire the target area. As the JTF stated it initially, one goal of the test was to determine the ability of a pilot to navigate from the initial point to the pop-up point and acquire the target area before reaching the wings-level point, at which the pilot stabilizes the aircraft and begins the dive toward the target area. Thus, the indicator that was to be sought was the range at which the pilot can acquire the target area. If there are no difficulties in finding the target area, this range should not be beyond the wings-level range. However, on four occasions, pilots did not find the target area before reaching wings-level. Thus, the JTF did not recognize potential operational difficulties in acquiring the target area.

The JTF also did not follow the analysis plan for addressing this objective. No analysis in the report of either the JTF or the System Planning Corporation examined how various initial-point departures affected success. No analysis in either report discussed the grouping of missions that was presented in the analysis plan. Although it was realized when the analysis plan was prepared that the sample size for many of the comparisons is small, so that it is difficult to be sure of statistical significances, valuable insights might have been gained from such analyses. For example, a comparison of target-area acquisition ranges for A-10 close air support missions flown under conditions of good visibility and poor visibility (with the time of day and the absolute

Figure 23
A-7 Pilot Learning: Target-Area Acquisition Range

humidity more or less equal) would have shown that poor visibility significantly reduced the target-area acquisition range (app. IV, item 10). The possibility that target-area acquisition ranges might be significantly shorter when visibility is poor should be of serious interest, given that the IIR Maverick was specially designed for such conditions.

In addition, no analysis was presented to show how the test pilots' increasing familiarity with the target area after repeated passes (known as the "effect of testing") was controlled for. When we examined the test data for record passes to determine what effect the testing itself had on the pilots' ability to acquire the target area, we found, for example, that was a favorite range at which the A-7 pilots acquired the target area during later passes (see figure 23). This suggests that the pilots' familiarity with the target area from repeated testing led to a test-specific ability to acquire the target area with no operational difficulties at this specific range. The A-10 pilots did not show this pattern of behavior.

Reporting problems

The JTF's reporting of the test results on the transition objective was, in some instances, unclear, unsupported, and misleading. No standards for evaluating target-area acquisition ranges were provided. The results for target-detection ranges were not supported by the data. The discussion of the pilots' ability to acquire targets with the missile omitted important facts.

The JTF stated that "long target area acquisition ranges are advantageous to the system operator because they permit more time to plan and execute the transition to the attack phase of the pass" (II.C.21, p. II-11). A median range of was reported for target-area acquisition, but no standards for judging the acceptability of this range were provided, even though some standards for the IIR Maverick's performance were documented and available at the time of the JT&E. For example, a system acquisition report prepared for the Congress specified a minimum launch range of for operations in poor weather. If the minimum standard for launch range were considered, a target-area acquisition range greater than in good weather would not be acceptable. However, the median range in the test was greater than

The JTF reported that for test passes the "target detection range was typically less than wings level range" (II.C.21, p. II-11). This suggests that it was not very difficult for pilots to locate targets once they had acquired the target area. However, the results the JTF reported on target detection ranges are not supported by the test data (app. II, item 11). The JTF's choice of the word "typically" implies "during most, if not all, passes," but the target detection range was less than the wings-level range during only 53 percent of all test passes for which data are available.

Finally, the JTF reported that

> "steering information from the A-7 INS [inertial navigation system], A-10 Pave Penny and visual cues from the realistic battlefield proved to be sufficient aids in placing the targets within the IIR Maverick field-of-view." (II.C.21, pp. II-7-8)

However, the JTF did not mention that neither the inertial navigation system nor the Pave Penny system alone was considered adequate for pinpointing the area or placing it within the IIR Maverick's field of view. Additional cueing was necessary. Thus, the JTF reported facts but did not fill in the details so that a clear and adequate conclusion could be drawn from them.

The System Planning Corporation summed up target acquisition this way:

> "By relying heavily on visual cues for timely target acquisition, the effective employment of the IIR Maverick system on the A-10 and A-7, or similar aircraft such as the F-16, may be limited to weather conditions and standoff ranges for which the pilot can visually observe the cues in the target area, and may require external information from a ground FAC [forward air controller], an airborne FAC or a pathfinder aircraft to orient the pilot to visual cues." (II.C.23, p. I-4, emphasis added)

This statement suggests that operational problems may be associated with transition in poor weather and at long standoff ranges. Unless these problems are acknowledged and overcome, the missile may be technically acceptable but of little value in poor weather. The JTF's failure to report the details of the pilot's ability to acquire targets is misleading.

Summary of threats to test results for the transition objective

In general, the test conditions we have discussed in this section probably led to overestimates of combat capability with respect to the transition objective. One important issue was not acknowledged in the definition of this objective--the ability to distinguish enemy from friendly forces. The implications of this omission were never discussed. In addition, the use of visual cues, the absence of navigational error, the small size of the target area, and the small number of "above average" test pilots make it difficult to generalize from the test results to combat. Changes related to the pilots' ability to find the target area and targets and to the test design matrix make it impossible to address some of the critical operational uncertainties of employing the IIR Maverick.

Besides the threats to the quality of the test because of the favorable test conditions, the JTF gave inadequate attention to analyzing and reporting some potential transition problems that were evident in the JT&E. The JTF's analysis did not fully examine the pilots' ability to find the target area or determine the effects of their becoming familiar with it. The analysis did not follow the proposed analysis plan. Some of the test results were unclear, unsupportable, and misleading in the way they were reported, and a more detailed analysis of the test data would have led to some useful information on the potential operational difficulties of the IIR Maverick.

Elaboration of test objective and reported results

One of the operational uncertainties in employing the IIR Maverick, a heat-seeking missile, is the pilots' ability to distinguish enemy tanks or armored personnel carriers from other sources of heat, called "thermal clutter," on a battlefield. Thermal clutter was simulated in the test with smoke, burning hulks representing previously struck armored vehicles, blank tank rounds, and flamethrowers. Vehicles and the equipment (a half-dozen vans, several generators, and a tent) necessary for the test's instruments surrounded the test area and made for additional sources of heat.

In the test, a pilot's ability to discern enemy targets was measured by the number of times (1) a pass was aborted from an inability to find the target area or a target, (2) an invalid target

VALID TARGET OBJECTIVE

JTF objective

Evaluate the IIR Maverick with respect to the "Capability of the operator to interpret a valid target in the presence of battlefield thermal clutter."

JTF conclusions

"The pilots were able to select valid targets from target arrays containing realistic thermal clutter. However, the JOT&E revealed the importance of proper ground training and practical experience in interpreting thermal signatures on the cockpit display. As the test progressed and the pilots gained experience, they were successful in using thermal signature contrast, shape, movement, and typical battlefield activity to discriminate valid targets."

(such as a burning hulk or tree) was chosen, and (3) a valid target (such as an enemy tank) was chosen. The test data show that pilots aborted passes percent of the time, chose invalid targets percent of the time, and chose valid targets percent of the time (app. IV, item 12). The JTF reported that a pilot's ability to distinguish valid targets from other heat sources on the battlefield depends on training and proficiency and added that two specific problems related to this ability. First,

> "The pilots stated that their largest problem was breaking out valid infrared signatures from the infrared signature of the surrounding terrain. This problem was related to the time of day. They had an easier time at night when there was more contrast between the armor and the relatively cool terrain." (II.C.21, p. II-34)

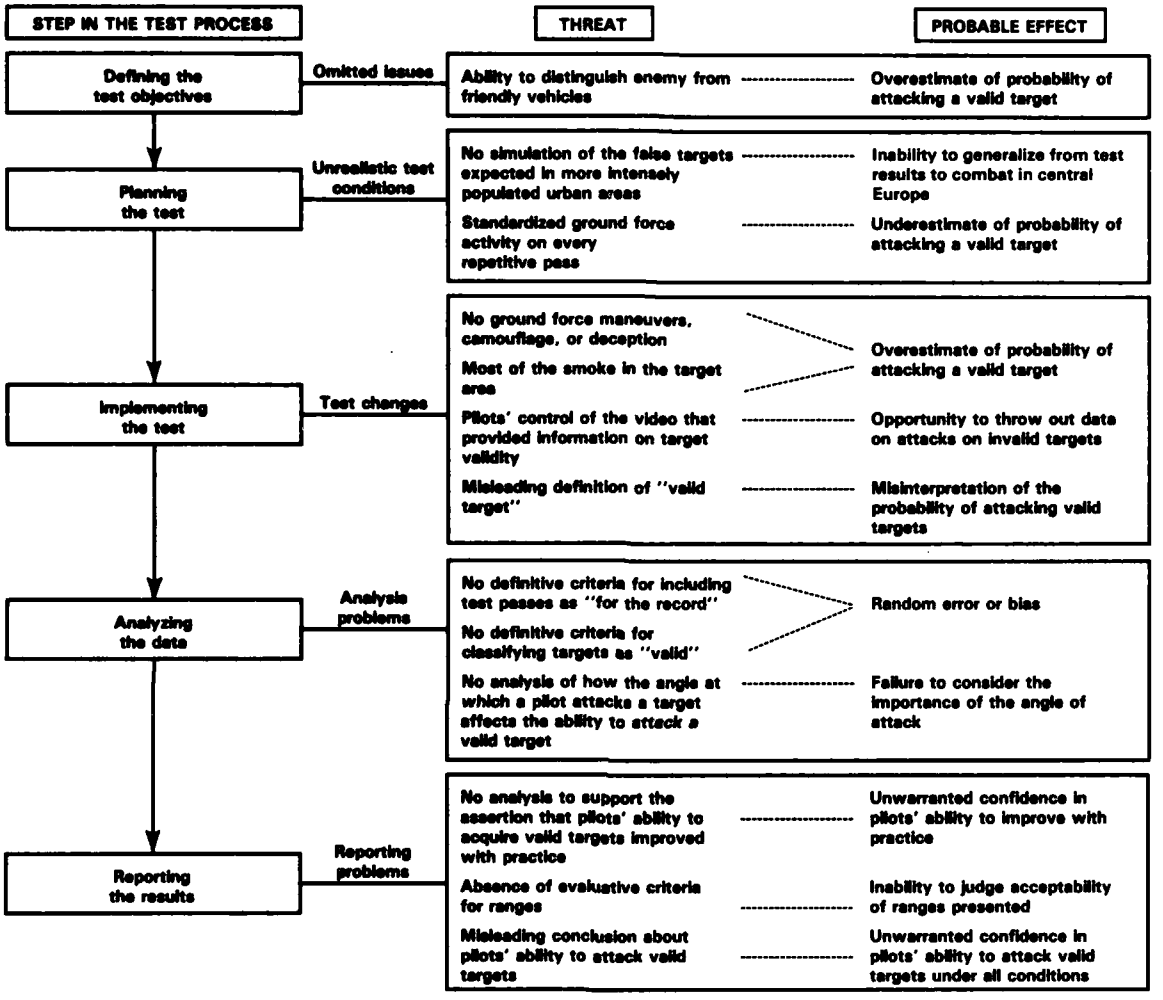
"Infrared signatures" refers to the temperature contrast between an object and its background.

Second,

> "The clutter presented by burning hulks and small grass fires appeared very bright and presented an irregular shape when viewed through the cockpit display. When the pilots did lock onto these fires, it was because the fires were either partially obscured by trees which tended to reduce their signatures while giving them a regular shape or the attack profile resulted in a shallow graze angle due to ceiling restrictions." (II.C.21, p. II-34)

"Graze angle" refers to the angle at which a pilot views the target. At low altitudes, which a pilot may fly because of poor weather and consequent ceiling restrictions, this angle becomes very small, so that only a small portion of the battlefield can be seen. Nevertheless, the JTF did not, in stating its conclusions, mention the problems it had identified in acquiring valid targets because of the time of day, the obstruction of trees, or shallow graze angles.

Figure 24
Threats to Test Quality: The Valid Target Objective



Threats to test quality

In figure 24, we show the major threats to the quality of test results for the valid target objective. There were threats at all five steps from defining the test objectives through reporting the results.

Omitted issues

As we mentioned in our discussion of the transition objective, friendly ground forces were not simulated in the test. Thus, the test did not examine or acknowledge the pilots' ability to discriminate between enemy and friendly vehicles. By omitting this issue, the test simplified the task of finding valid targets on the battlefield.

Unrealistic test conditions

Two test conditions make it difficult to generalize from the test results to combat in central Europe. One has to do with the failure to use on the simulated battlefield the types of false target that, according to Air Force documentation, are to be expected in the densely populated, urban areas of Europe. Moreover, the test target area had only one type of soil (a mixture of loam, sand, and clay) and only one type of tree (southern pine). Taken together, these limiting factors make it impossible to generalize from this test to the diverse environment in Europe, even though the purpose of the test was to address the operational uncertainties of employing the IIR Maverick in a midintensity conflict in central Europe.

The other test condition is that the ground force activity on each repetitive test pass was standardized, and the specificity of this activity unrealistically aided the pilots in finding valid targets. In the close air support scenario, two or three blank rounds were fired by each of one to eight tanks during the target acquisition phase of each pass. The pilots used these gunfirings as cues in finding valid targets. In combat, gunfirings would be coming from both friendly and enemy forces and would not be timed to occur during target acquisition.

For every test pass in the preplanned interdiction scenario, the tank convoy always used the same road and it was always lit with burning hulks that were always located at the same six predetermined checkpoints. This target array gave the pilots the opportunity to learn very quickly which were the valid targets and which were only the burning hulks. Evidence in the test documentation shows that some of the pilots could discern "burning hulks" that had not been lit, because they had become so familiar with the target.

Test changes

One component of the test plan that was not implemented as proposed and three test events that were not planned for may have affected the quality of the data on valid target acquisition. The test plan called for reasonable and prudent simulation of an enemy counter to IIR Maverick by ground force maneuvers, camouflage, and deception, but the target array did not react at all to the air attack. This made it easier for pilots to find valid targets.

Flares, smoke pots, and fires of diesel fuel were ignited randomly up to 2 to 3 kilometers from the target area, but most of the battlefield smoke usually arose within the target area. This helped direct pilots to the target area more quickly than might happen in combat if there were several areas of smoke. It also allowed the pilots more time than they might have in combat to find valid targets before becoming vulnerable to enemy air defenses.

In the test, the pilots controlled the video system that provided the data for determining a target's validity. Sometimes data must be collected by the persons whose performance is being measured (the pilots, in this test), but it is difficult in such instances to insure the quality of the data. The fact that 30 percent of the not-for-the-record passes were the result of a failure of the video system suggests that the pilots had opportunity to discard invalid target data in order to improve their test performance. The video should have been carefully examined during the test's implementation to make sure that the failures were not induced by the pilots. The JTF report provides no explanation for the video failures.

A valid target attack required that the missile seeker be locked on to a tracked armored vehicle before the pilot declared it launched. For some passes, however, the seeker stopped following a target immediately before the pilot gave the launch call, and the pilot was credited with acquiring a valid target anyway. Consequently, the reported probability of attacking valid targets may be optimistic.

Analysis problems

Analysis problems included the failure to specify definitive criteria for "for-the-record" passes and "valid" targets. No analysis was made of the effect on the data on valid target acquisition of the angle at which pilots attacked targets.

The rates at which pilots acquired valid targets are presented discrepantly in various reports on this test, and most of the discrepancies can be attributed to the fact that "for-the-record" passes were not counted systematically (app. IV, items 13 and 14). This indicates that there were no definitive criteria for classifying passes as "for the record." The reports also reveal a discrepancy in the classification of targets as "valid," indicating that the criteria for classifying targets as valid were also not definitive. The credibility of the test results is doubtful.

The JTF noted that there were many times in the test when ceiling restrictions led the pilots to sacrifice the angle at which they could view a target, and the JTF noted that the pilots sometimes acquired invalid targets when the angle of attack was so shallow that they could see only a small portion of the target area. Although these angles were recorded for every test pass, no systematic analysis was made of their effect on the ability to attack valid targets. If ceiling restrictions make shallow angles of attack necessary, as may happen in poor weather, then the IIR Maverick's operational utility may be limited in poor weather. The importance of this was overlooked in the JTF's analysis.

Reporting problems

The JTF's report of some test results on this objective was not supported by the test data, provided no standards for evaluat-

ing launch ranges, and supplied an unwarranted conclusion on the pilots' ability to attack valid targets.

The JTF reported that learning occurred during the test, which implies that the pilots became more proficient in attacking valid targets as the test progressed, but the JTF conducted no analysis of the effect of learning on the success of finding valid targets. When we examined the data, we found no continuous learning curve in the pilots' ability to attack valid targets. Our analysis of the test data for the relation of learning to valid target acquisition is summarized in figure 25 (on the next page). We found that some learning may have occurred when the time of day is controlled for but that, otherwise, the JTF's report places undue confidence in the pilots' ability to improve with practice.

The JTF reported its analysis of launch ranges, showing distributions, means, and medians, but provided no evaluative criteria with which to judge the adequacy of these ranges. When we reviewed the JTF's analysis, we found that only 4 of the 22 test missions that were reported had average launch ranges for valid targets that were greater than (app. IV, item 15). However, earlier standards for the IIR Maverick's performance that had been established and documented before the JT&E--the system acquisition report prepared for the Congress is one example--state a minimum launch range of in poor weather. Had this standard of system performance been considered, the launch ranges in the test would not have been acceptable.

The JTF concluded that the test pilots were successful in acquiring valid targets and, in doing so, omitted referring to many important details about potential operational difficulties that it had presented in its report. For example, the JTF stated that ground visibility of 3 miles or less gives a probability of attacking a valid target and that the probability increases to when ground visibility is 4 miles or more, even though the JTF noted that the sample size was not large enough to establish a statistically significant association. The JTF also stated that

> "the probability of attacking a valid target is dependent to a significant degree upon surface wind speed. Specifically, the probability of attacking a valid target appears to be increased when surface winds are below 5 knots." (II.C.21, p. II-24)

The JTF stated that "during the test, the pilots were rarely close enough at launch to classify a target as a tracked vehicle from passive thermal features alone," concluding nevertheless that the pilots were "successful" in using the temperature contrast between the target and its background (II.C.21, p. II-25). Overall, it appears that the JTF's conclusion that pilots can successfully attack valid targets under all conditions is not based on the JTF's own critical report of the operational feasibility of the IIR Maverick under specified conditions.

Figure 25
The Probability of a Pilot's Acquiring
a Valid Target with the IIR Maverick

Summary of threats to test results for the valid target objective

In general, the favorable test conditions and questionable test procedures that we have discussed in this section probably led to an overestimation of the probability of attacking valid targets and throw the quality of the test results into doubt. Although the IIR Maverick was tested in close air support, the pilots' ability to distinguish enemy from friendly vehicles was not acknowledged or tested, because no friendly ground forces were simulated. The test did not simulate the false targets that can be expected in central Europe, so that one cannot generalize from the test results to combat in central Europe. The standardized activity, the absence of enemy maneuvers, camouflage, and deception and the fact that most of the smoke arose in the target area all made the task of finding valid targets easier than can be expected in combat. The pilots' control over the video system providing them with data on target validity gave them opportunity to omit invalid targets from their operations. The definition of target validity failed to account for the fact that the missile's eventual target may be different from the target a pilot selects at the time of launching the missile.

In addition, the problems in the way the test results were analyzed and reported that we have discussed in this section detract from their usefulness. The criteria that the JTF used to classify test passes and targets were not definitive, as evidenced by the several different classifications that various reporting sources presented. The JTF suggested that attack angles may be important but conducted no analysis on the question. We found, contrary to the findings reported by the JTF, that practice did not necessarily improve the pilots' ability to attack valid targets, the reported launch ranges are generally unacceptable in comparison with documented standards for the IIR Maverick's performance, and the conclusion on the pilots' ability to attack valid targets under all conditions is unwarranted given the operational difficulties that were evident in the employment of the IIR Maverick.

Elaboration of test objective and reported results

Before the JT&E, the requirement for cueing aids was an operational uncertainty for the IIR Maverick. In other words, it was not known what cues, if any, a pilot would need to find the target area or to find valid targets. The options for cueing aids were an inertial navigation system, Pave Penny or Pave Tack, communication from forward air controllers, and visual aids. Since the IIR Maverick is intended for use at night and in poor weather, visual cues alone might not suffice.

The JTF's only analysis of cueing aids was presented in an analysis of Pave Penny missions with A-10 aircraft (app. IV, item 16):

"While the observed probability of attacking a valid target was greater without the Pave Penny system without Pave Penny versus with Pave Penny), statistical tests for dependency between use of Pave Penny and pass outcome reveal that the observed differences are not significant." (II. C.21, p. II-30)

The JTF's conclusions on cueing are somewhat confusing because they are contradictory. On the one hand, the JTF concluded that cueing in the test was sufficient (II.C.21, pp. II-8 and II-35); on the other hand, it concluded that cueing is not only useful but also essential (II.C.21, p. II-29). The JTF reported that it may be necessary to add some cueing aids to the cockpit if the IIR Maverick is to be employed successfully when visual cues are scarce and also reported that adding cueing aids to the cockpit is not necessary. The JTF concluded that Pave Penny is valuable as a cueing aid and reported that overall performance was poorer with the Pave Penny than without it.

CUEING OBJECTIVE

JTF objective

Evaluate the IIR Maverick with respect to "The requirement for cueing."

JTF conclusions

"Closely related to the issue of transition, the JOT&E demonstrated that current tactics, procedures onboard navigation systems, and visual battlefield activity provide sufficient cueing information for target area acquisition and target detection. Given target coordinates, the A-7 [inertial navigation system] provided accurate steering to the target area, thus suggesting the need and utility of a similar system for the A-10. The A-10 Pave Penny provided accurate cueing to the target arrays and was particularly valuable during the test missions when visibility was reduced in blowing dust."

"the IIR Maverick is not suitable as a target search device unless the search is small or the environment is target-rich. Cueing to the target area is therefore essential to IIR Maverick success."

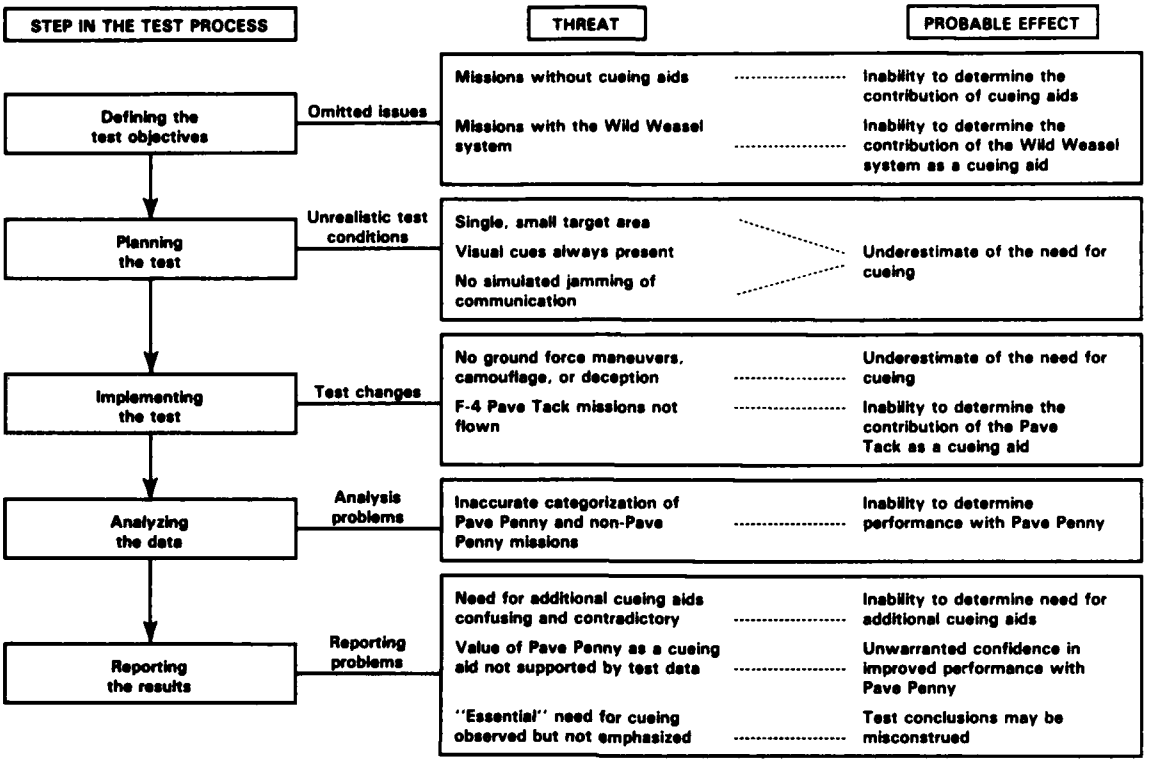
"In a low intensity environment where visual cues are scarce, additional cueing aids, such as a precise onboard navigation system or a laser spot seeker, would be a useful enhancement to the target acquisition task."

"the missile can be succesfully employed by the A-10 or A-7 without adding other cueing aids in the cockpit."

Threats to test quality

Figure 26 lists the major threats to the quality of the test results with respect to the cueing objective. There were threats at all five steps from defining the objectives through reporting the results.

Figure 26
Threats to Test Quality: The Cueing Objective



Omitted issues

Although the Air Force has stated that the IIR Maverick system can be operated without acquisition aids, that mode of use was not included in the definition of the test's objectives. Consequently, the contribution of cueing aids could not be determined because no missions were conducted without cueing aids. In addition, the F-4G, called the "Wild Weasel" system, a sensor for locating radar emissions, had been specified for use in poor weather but was not tested as a cueing aid for the IIR Maverick.

Unrealistic test conditions

Three test conditions make it difficult to generalize from the test results to combat: the one small target area, the presence of visual cues, and uninterrupted communication. The small size of the target area probably made the four pilots overly familiar with it, diminishing their need for cues. Thus, the need for cueing may have been underestimated. Visual cues were present during all test missions. The test design did not specify that any missions be flown without visual cues, despite the fact that the IIR Maverick is intended for use when there may be no visual

cues. Consequently, the test results probably underestimate the need for cueing in the type of weather that makes visual cues sparse or nonexistent. The test did not simulate enemy jamming of the communication between the pilots and the forward air controller, who provided target cueing information to the pilots. Since the pilots received complete, uninterrupted information, they may have depended more on it than on other cues they might use in combat in which enemy countermeasures result in sporadic information on the location of targets. Therefore, the test results probably underestimate the need for cueing.

Test changes

The test plan recommended that the enemy ground forces use maneuvers, camouflage, and deception in reaction to air attack, but this plan was not carried out. Consequently, the test data do not show what cues pilots would need if enemy ground forces were responding to attack in actual combat. In addition, the test plan called for some missions with F-4 aircraft using the Pave Tack cueing aid, but none were conducted, a decision that the Director for Research and Engineering made. According to the JTF's final report, the reasons that were given for deleting the F-4 missions were cost and success in employing the IIR Maverick with the A-7 and A-10, as demonstrated in the JT&E. However, the change means that the Pave Tack was not tested as a cueing aid for the IIR Maverick.

Analysis problems

The JTF's sole analysis of cueing aids, on the Pave Penny, was flawed by the way passes were categorized. On some passes that were categorized in the analysis as Pave Penny passes, the pilots had chosen not to use the Pave Penny as a cueing aid for target acquisition, even though it was available. Consequently, a comparison of passes with and without the Pave Penny as the JTF categorized them can consider only the availability of the Pave Penny, not its performance as a cueing aid.

Reporting problems

The JTF's conclusions are contradictory on the need for cueing and the value of the Pave Penny as a cueing aid. The JTF concluded that the IIR Maverick can be successfully employed without cueing aids, having also reported that cueing aids may be useful when visual cues are scarce. Because visual cues were always present in the test, it is not possible to determine from the test results the need pilots may have for additional cueing aids. Further, the JTF reported, in an analysis that was flawed (as we discussed above), that the IIR Maverick's performance with the Pave Penny as a cueing aid was poorer than without it. Yet the JTF also concluded that the Pave Penny was valuable, which suggests that performance was better with it. Finally, the JTF had observed that cueing to the target is essential to the IIR Maverick's success but failed to incorporate this observation in

the overall summary. Because this information was not fully reported, the importance of cueing in the use of the IIR Maverick may be misconstrued.

Summary of threats to test results for the cueing objective

Various aspects of the test make it difficult to fully assess the requirement for cueing. The one small target area, the presence of visual cues and uninterrupted communication, and the lack of response from enemy ground forces to air attack means that test estimates of the need for cueing may be too low. The failure to conduct missions (1) with no cueing aids, (2) with the Wild Weasel as a cueing aid, and (3) with the Pave Tack as a cueing aid makes it impossible to compare proposed cueing requirements. In particular, the failure to conduct missions without cueing aids in the test led to the omission of baseline data.

The JTF analyzed only the Pave Penny as a cueing aid, and the analysis was flawed. The JTF's conclusion on the value of the Pave Penny places unwarranted confidence in its usefulness, while the need for additional cueing aids, as reported by the JTF, cannot be supported by the test data. The JTF's summary conclusion on the usefulness of cueing aids may be misconstrued because of the JTF's failure to incorporate its own observation that cueing is not only useful but also essential.

Elaboration of test objective and reported results

The utility of the IIR Maverick system is also based upon the pilot's ability to find and attack the enemy and survive these tasks. The probability of surviving in the presence of enemy air defenses (such as surface-to-air missiles and antiaircraft guns) was an operational uncertainty about the IIR Maverick. Nevertheless, enemy air defense action was not simulated in the IIR Maverick test, although data were collected on the extent of exposure of the test aircraft to enemy air defenses.

SURVIVABILITY OBJECTIVE

JTF objective

Evaluate the IIR Maverick with respect to the "Extent of exposure to forward-area defenses while accomplishing the functions needed in the delivery of the IIR Maverick missile."

JTF conclusions

"The long standoff ranges (compared to existing inventory weapons) achieved with the IIR Maverick enhance the survivability of delivery aircraft against enemy ground defenses. The period of greatest vulnerability, that is the time the aircraft were wings level after pop-up until simulated launch or abort, was recorded on all passes. These wings-level times and standoff range data from the A-10 night missions were used for a survivability analysis by HQ USAF Studies and Analysis. The results of their analysis are published separately in annex C (Secret) of this report."

That is, data were collected on wings-level time (the point of aircraft stabilization as the pilot prepares to dive toward the target area to launch the missile or abort the mission) and on standoff, or launch, ranges. Wings-level time, also known as "tracking" time, is important because it is a period of exposure in which the aircraft is more likely to be attacked by enemy air defenses the longer it grows. Launch range is important because it is a distance at which the aircraft is less likely to be attacked by enemy air defenses the farther away it is from the target, given that these defenses are usually close to the target. The JTF reported the distribution of launch ranges for all passes and the distribution of tracking time for valid, invalid, and aborted passes (app. IV, items 17 and 18).

The JTF did not specifically report results on survivability, but the Air Force Studies and Analyses group did (in an annex to the JTF report). It reported that,

The System Planning Corporation conducted a survivability analysis, reporting that 66 percent of the test launches would be subject to interception by the Soviet mobile heat-seeking surface-to-air missile, the SA-9, and that

(II.C.23, p. IV-10)

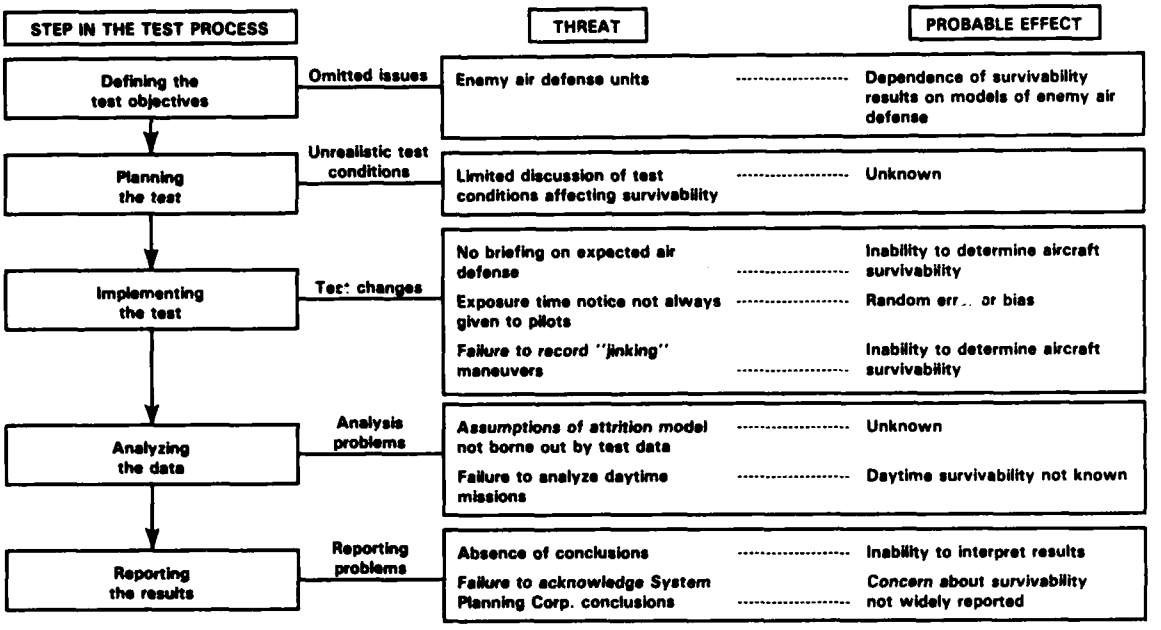
Threats to test quality

Figure 27 shows the major threats to the quality of the test results with respect to the survivability objective. The test quality was threatened at all five steps from defining the objectives through reporting the results.

Omitted issues

Armored personnel carriers simulated unspecified air defense units in the ground scenario, but enemy air defense action was not played in the test. As a result, survivability could be estimated only from computer models of expected enemy activity. This meant that the test yielded data on and made it possible to determine the wings-level time and the launch ranges of the friendly aircraft, while it yielded no data on and made it impossible to determine how often enemy air defense units might be able to detect, acquire, and attack friendly aircraft. The implications of using a computer model to simulate enemy action are discussed in the section below on analysis problems for the survivability objective.

Figure 27
Threats to Test Quality: The Survivability Objective



Unrealistic test conditions

In comparison with the detailed plans for the other objectives we have discussed, there was very little on this objective in the test plan. The test plan stated that

> "this analysis will examine the time and evasive maneuvers taken by the aircraft up to the point of launch or pass abort. Based on the scenario locations of air defenses, these data and previous studies of the IIR Maverick, will be used to make a partial assessment of survivability." (II.C. 13, p. 10)

According to the test plan's "intelligence/threat scenario," the enemy air defenses in a close air support mission would be the Soviet antiaircraft gun ZSU-23-4, the surface-to-air missiles SA-7 and SA-9, and either the SA-6 or the SA-8; in a preplanned interdiction mission, they would be the same except that the SA-8 would not be used. According to the test plan, the pilots were to be briefed before each mission on the expected air defenses and their probable locations and they were to be instructed to use tactics for a minimum of exposure time to enemy threats. The forward air controller was to give the pilots an apparently arbitrary time limit of after reaching wings-level, presumably the maximum time the aircraft could be exposed without being attacked by enemy air defenses.

Test changes

There is no indication that the aircrews were briefed as planned on the expected enemy air defenses. The tactics they flew were, therefore, not chosen in response to a specific threat to their survival. Thus, the test data on wings-level time and on launch ranges cannot be used to determine the ability of friendly aircraft to survive enemy air attack. Moreover, the pilots' debriefing forms indicate that the forward air controller did not always give the pilots the warning that was planned. An examination of the test data reveals, indeed, that of the passes that resulted in an attack on a valid target exceeded Thus, not all test passes represented the same survivability tactics. Finally, tactical maneuvers to take quick evasive turns to avoid enemy air defenses, called "jinking," if they occurred, were not recorded. It is difficult to assess the survivability of aircraft if nothing is known about these maneuvers.

Analysis problems

The JTF reported that the "long standoff ranges" improved survivability but presented no supporting data. Launch ranges were reported, but no evaluative criteria for defining exactly what constitutes a "long standoff range" were provided. The JTF stated that the IIR Maverick launch ranges were long, in comparison with weapons in the existing inventory, but gave no data to support this comparison.

The Air Force Studies and Analyses group used the basic flight profiles for A-10 night missions in the JT&E to determine aircraft survivability from an attrition model, but the model was based on many assumptions that differed from the test conditions. For example, no A-10 in the JT&E carried outside devices for use against enemy air defenses, but the attrition model assumed that all did. Similarly, the analysis considered only night missions, so that it was not possible to report on daytime survivability as intended. The reason given for using only night missions was that the altitudes of the daytime missions that were flown did not meet the standards for tactics that had recently been developed for entering a combat area in daytime at very low altitude.

Reporting problems

The Air Force Studies and Analyses group reported on survivability as expected attrition per nighttime A-10 pass but did not interpret the figures and provided no conclusions on the acceptability of the attrition results. It is impossible, for example, to determine whether losing A-10's to enemy air defenses (in particular, during night missions aided by visual cues is good, bad, or indifferent.

Part of the System Planning Corporation's analysis of survivability was based on the launch ranges of the expected enemy air

defense units. The conclusion was that the

if aircraft carrying the IIR Maverick were to survive the JT&E passes. Although this conclusion suggests that survival may be problematic, this potential problem was not reflected in the discussions in the JTF or Air Force reports.

Summary of threats to test results for the survivability objective

The IIR Maverick JT&E was very limited in how it addressed survivability. No enemy air defense unit action was simulated, and survivability analysis depended on models of the purported capability of enemy air defenses, not on empirical data about their capability. There was very little planning with regard to the survivability objective, and the few plans that were specified were not followed: the pilots were not briefed on the expected enemy air defenses, they were not always warned of the exposure limit, and the time periods of maneuvers they made in response to enemy threats, if they took any, were not recorded. As a result, it is highly questionable whether the survival maneuvers they flew in the JT&E are representative of flight in response to an actual enemy threat.

Despite these limitations in the JT&E data, survivability was addressed. The JTF did report standoff ranges but gave no criteria for judging their acceptability. The Air Force did analyze A-10 night missions but with a model based on assumptions that the test data did not meet, and it reported an attrition rate but without interpretations. The System Planning Corporation concluded that survivability may be a problem for the IIR Maverick, given the test data, but this conclusion was not acknowledged by either the JTF or the Air Force.

Elaboration of test objective and reported results

One concern that the OSD expressed before the JT&E began was whether operator workload in single-seat aircraft diminishes the IIR Maverick's effectiveness compared with workload in a two-seat aircraft carrying both a pilot and a navigator. The many steps that are required in using this missile may be difficult to accomplish in a single-seat aircraft by a pilot who must navigate, find the target area, attack the enemy, and successfully avoid enemy attack all alone. Therefore, the OSD requested comparative data on the IIR Maverick's effectiveness in two-seat aircraft. However, all 24 missions in the JT&E were flown with single-seat A-7 or A-10 aircraft. The original plan to conduct missions in the two-seat F-4 was never carried out.

The JTF concluded that workload was not a problem in employing the IIR Maverick in single-seat aircraft, nevertheless recommending four ways to reduce workload: (1) using an automatic radar warning system to place expected threats in an order of priority

SINGLE-SEAT AIRCRAFT OBJECTIVE

JTF objective

According to the test plan, this objective was to assess the IIR Maverick employment capabilities in single-seat and two-seat aircraft operations. The JTF final report stated that the objective was to evaluate the IIR Maverick with respect to the "Capability of accomplishing the IIR Maverick delivery function in a single-place aircraft under operational conditions."

JTF conclusions

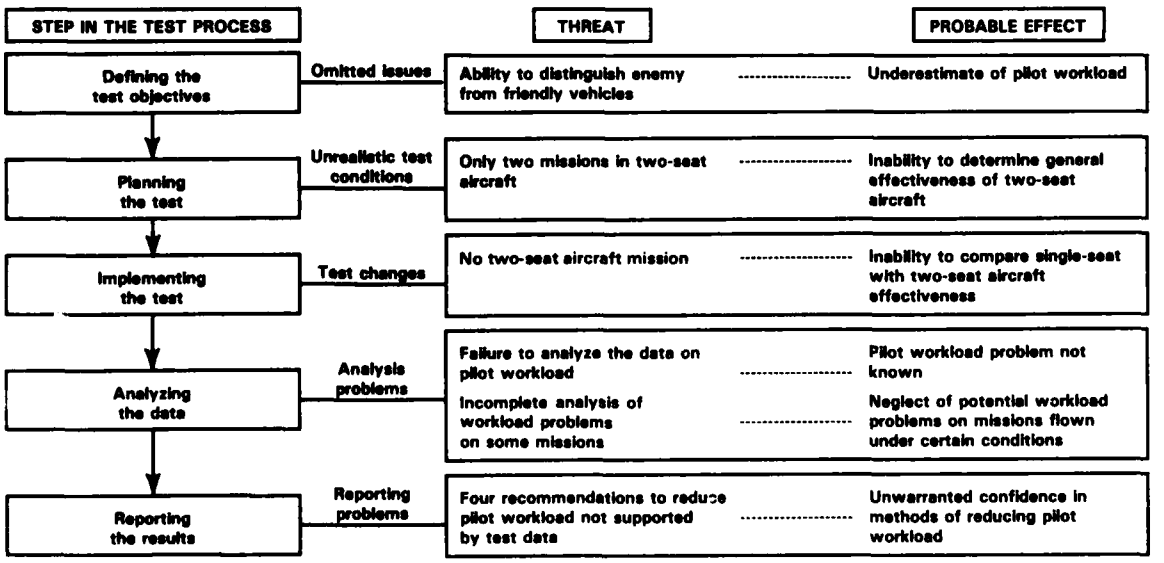
"The pilots participating in the test reported few problems in performing the IIR Maverick attack mission in single-seat aircraft. Single-seat employment was successful both day and night and in limited visibility conditions of rain, fog, haze and blowing dust and heavy battlefield smoke. Based on the recommendation of the JTF citing the ease of single-seat employment, [the Director for Defense Research and Engineering] concurred in the cancellation of a series of dual-seat F-4 test missions. Therefore, comparative data to determine dual-seat employment advantages, if any, were not obtained."

for the pilots; (2) placing the target designated by the Pave Penny automatically within the IIR Maverick's view; (3) reducing the size of the tracking gates, increasing the sensitivity of the seeker, and replacing the edge tracker with a centroid tracker; (4) using the dual-field-of-view seeker to enable pilots to see more of the target area.

Threats to test quality

Figure 28 shows the major threats to the validity of the test results with respect to the single-seat aircraft objective.

Figure 28
Threats to Test Quality: The Single-Seat Aircraft Objective



There were threats at all the steps but, because this objective was tested in such a limited manner, we present only summary observations about the results.

Summary of threats to test results for the single-seat aircraft objective

The omission of an important feature of combat in close air support--distinguishing enemy from friendly vehicles on the battlefield--reduced the test pilots' workload. This portrayal of combat was unrealistic, and the data may underestimate pilot workload with the IIR Maverick.

According to the test plan, one single-seat aircraft mission in an A-10 was to be compared with one two-seat aircraft mission in an F-4. No explanation was provided for these small numbers. After the preliminary results of the A-7 and A-10 missions were reported to the Director for Defense Research and Engineering, the F-4 missions were dropped from the test because of the "ease" with which the A-7 and A-10 pilots used the IIR Maverick in the JT&E missions already conducted and because of the "substantial" costs of moving the test to another location that would accommodate the F-4. Consequently, no comparison of single-seat and two-seat aircraft performance was made.

The JTF not only did not compare workloads in aircraft with one- and two-member crews but also did not conduct any specific analysis of test data on pilot workload in the single-seat aircraft. The JTF reported that "the pilot was able to employ the missile even though results varied, depending on actual test conditions," but did not analyze how the results that varied were related to workload (II.C.21, pp. II-31-35). Data on workload were available from the pilots, as is evident from the System Planning Corporation's analysis:

> "Debriefing of the pilots indicated no specific workload problem except for turbulence on several passes. . . . In only three missions during the test, all with the A-10, was a workload problem indicated, and these were the result of turbulence and buffeting. . . . However the test did not place the pilots under the stresses that they would encounter in combat." (II.C.23, pp. I-8 and III-38)

The three missions for which the pilots said workload was a problem were the only A-10 missions flown on totally cloudy days with relatively high humidity. These two weather conditions--cloud cover and high humidity--were cited in the test plan as the "two main weather" factors that "affect the operational success" of the IIR Maverick. Further examination of the pilots' debriefing forms, however, reveals that low dive angles and reduced infrared signatures would also contribute to workload problems under various combat conditions. These problems went unnoticed because the JTF did not conduct a detailed analysis of the test data.

The four recommendations that the JTF made for reducing pilot workload appear to contradict its assertion about the "ease" of single-seat employment. No test data were given to support the recommendations, so that any confidence that may be placed in them as ways of reducing pilot workload is unwarranted.

In summary, the JTF did not compare single-seat with two-seat aircraft performance in the employment of the IIR Maverick. The favorable test conditions may have made it easier to employ the IIR Maverick in the test than in combat. The JTF failed to analyze the test data to determine what, if any, test conditions resulted in workload problems for the pilots of the single-seat aircraft. Total cloud cover and high humidity, among other things, may create workload problems, but the JTF ignored them in the analysis. The JTF's recommendations for reducing pilot workload are, thus, unsupported by the JT&E and contradict the JTF's conclusions.

Elaboration of test objective and reported results

Countermeasures by an enemy either are intended to prevent a weapon from working well or may inadvertently thwart its use. In the IIR Maverick JT&E, the simulation of intentional countermeasures consisted of

Inadvertent countermeasures consisted of

In preplanned interdiction, however, only bonfires and burning hulks (both intended as checkpoints) could be considered inadvertent countermeasures, and no intentional countermeasures were used.

In each of the 18 close air support missions, the first pass included (these were considered intentional) and the second included (these were considered inadvertent or unintentional). The pilots typically encountered all the inadvertent countermeasures in the close air support passes. In

COUNTERMEASURES OBJECTIVE

JTF objective

Evaluate the IIR Maverick with respect to the "Degradation of system utility by the use of countermeasures."

JTF conclusions

"Both inadvertent and deliberate countermeasures were employed during the JOT&E. A detailed analysis of the effects of countermeasures, deliberate and inadvertent, will be performed by the U.S. Army Office of the Test Director, Joint Services Electro-Optical Guided Weapons Countermeasures Test Program and will be published separately as annex B (Secret) of this report."

addition,

Details of the
attempted during the close air support missions are presented in appendix IV, item 19.

The JTF and the System Planning Corporation both reported that passes were "for-the-record" passes and included (a target whose validity could not be determined). The "conclusions" by the JTF that we have quoted in the accompanying display are not, in fact, conclusions, but the JTF presented them as its assessment of and conclusion on this objective. In contrast to the JTF and SPC reports, the detailed report by the Joint Services Electro-Optical Guided Weapons Countermeasures Test Program stated that resulted in "for-the-record" passes.

Threats to test quality

Figure 29 on the next page presents the major threats to the quality of the test results as they relate to the countermeasures objective. There were no significant threats from the test-planning step.

Omitted issues and unrealistic test conditions

A 1975 static test on the IIR guidance unit reported that the unit failed to maintain its lock "Only when

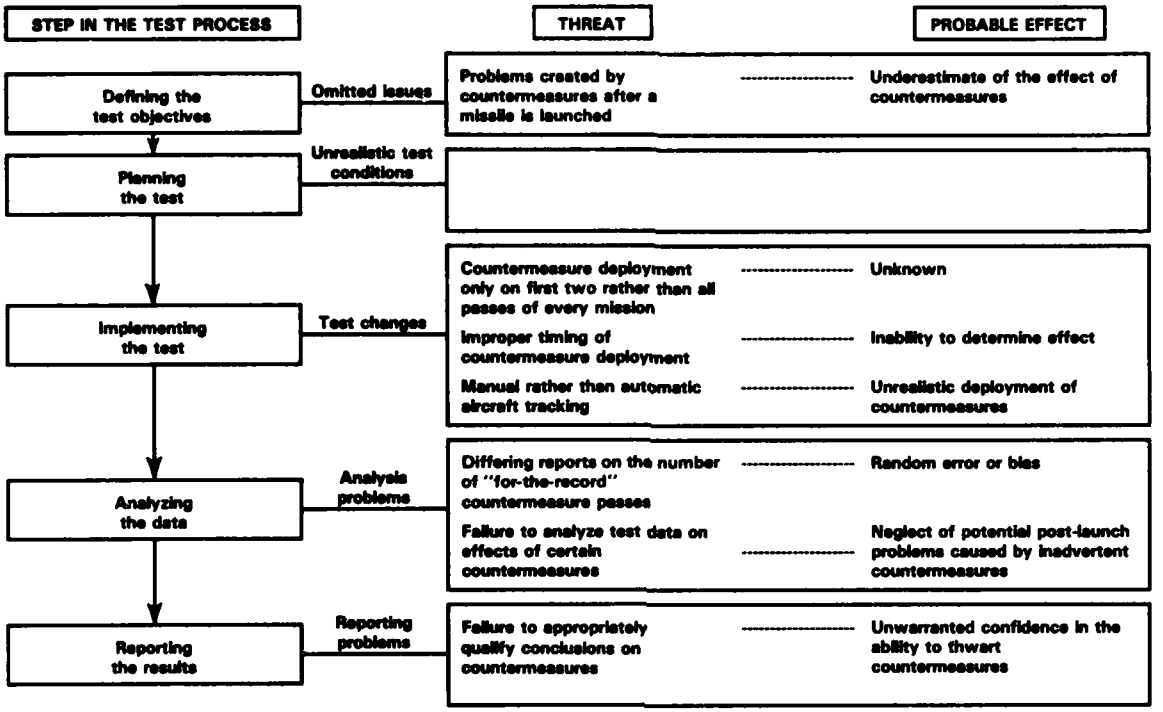
Nevertheless, the JT&E was designed to examine the IIR Maverick's operation only

. The missile was tested, however, for its susceptibility to , testing that had been recommended as a result of the 1975 static test.

Test changes

The test plan called originally for countermeasures on every pass, but was used on only the first two passes of every mission. Consequently, the first two passes of every mission are different from the rest. No explanation was given for

Figure 29
Threats to Test Quality: The Countermeasures Objective



the change in the test reports. The Electro-Optical Guided Weapons Countermeasures group reported that

" (II.C.19, p. 11)

Consequently, the effects of on the operation of the IIR Maverick could not be determined.

Automatic tracking of the aircraft was planned, but the unavailability of automatic equipment made manual operations necessary.

Analysis problems and reporting problems

The differences we noted earlier in the numbers of countermeasures tests that were reported indicate that the criteria for

counting passes for the record were not definitive. This lessens the credibility of the results. Moreover, only the System Planning Corporation reported that

While such problems were not to be specifically addressed in the test, the JT&E results did suggest that

These problems were not noted in the JTF report. The failure to recognize this led to a report of unwarranted confidence in the ability of the IIR Maverick to defeat countermeasures. While reporting that countermeasures had no effect on the IIR Maverick, the System Planning Corporation did qualify this assertion by describing the constraints on the test that seriously hampered the ability to address the effects of countermeasures.

Summary of threats to test results for the countermeasures objective

In general, the most serious threats to the quality of results for the countermeasures objective were implementation constraints. Countermeasures were not implemented as they would be in combat, and the result does not give a realistic estimate of how countermeasures may affect the IIR Maverick's utility.

Elaboration of test objective and reported results

Two of the original operational uncertainties of the IIR Maverick system involved the extent to which the thermal characteristics of targets and the battlefield can affect a pilot's ability to discern enemy targets. The infrared seeker senses minute differences in temperature, which are run through a mechanical scanning system and displayed on a TV-like monitor in the aircraft's cockpit. Since all objects emit heat, the use of the IIR Maverick requires that all battlefield sources of heat be evaluated in relation to the heat emitted by enemy targets. That is, distinguishing the thermal signatures of enemy vehicles from the thermal signatures of other things on the battlefield is critical to using the IIR Maverick successfully as a weapon.

THERMAL CHARACTER OBJECTIVE

JTF objective

Evaluate the IIR Maverick with respect to "The thermal character of the proposed targets" and "The thermal character of the battlefield."

JTF conclusions

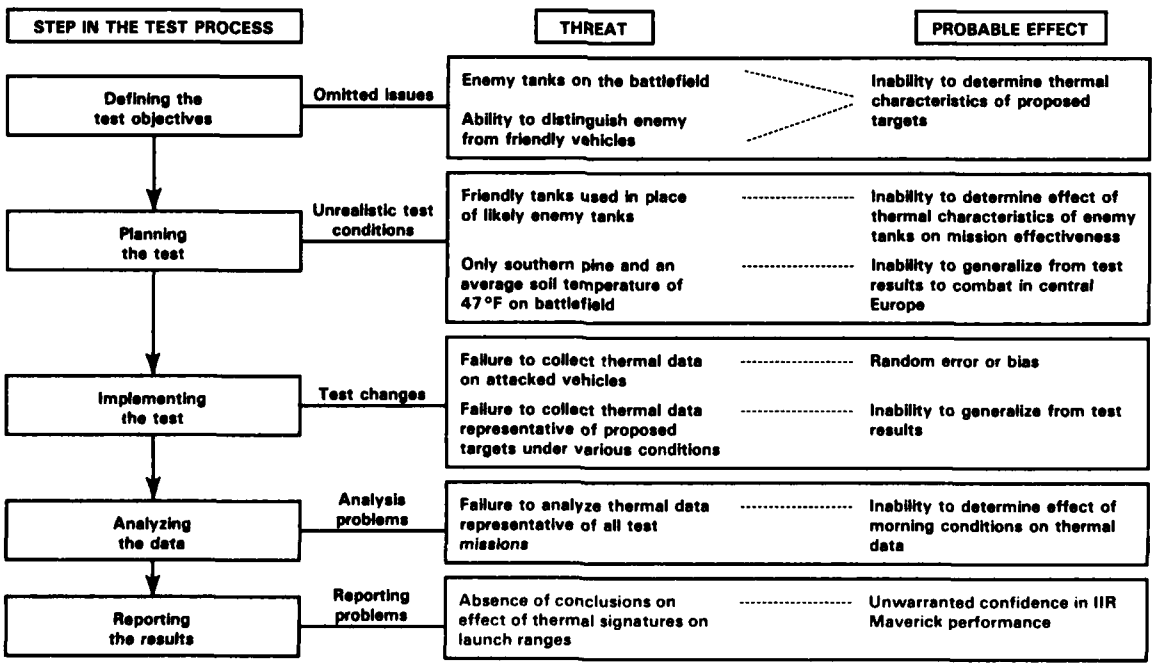
"Extensive data on the thermal character of the proposed targets and the battlefield were collected from both ground-based and airborne sources. This included a follow-on effort by the [Naval Weapons Center] with their S-3 FLIR-equipped A-6 aircraft at Ft. Sill, Oklahoma on 20 April 1977 which obtained thermal imagery of 155 howitzer and 8 inch live artillery fire. All thermal data were provided to System Planning Corporation and HQ USAF Studies and Analysis for their independent analyses."

Two types of thermal measurement of target vehicles were collected in the JT&E--a simple measurement with a precision radiation thermometer (a PRT-5) of apparent temperatures relative to a bare earth background and a more complex measurement of radiant temperature based on an analysis of calibrated thermal imagery. However, the JTF reported only the simpler of the two, giving the differences between tank targets and their backgrounds with the associated launch ranges (app. IV, item 20). After its analysis of these data, the JTF stated that "there is a statistically significant increase in launch range associated with higher thermal contrast" (II.C.21, p. II-28). In other words, the greater contrasts in temperature between the target and the background led to the launching of the missile at longer ranges.

The System Planning Corporation analyzed the signatures produced by calibrated thermal imagery and reported that when thermal contrasts were high, approximately of the targets that the pilots selected were valid, regardless of the launch range. When the contrasts were low,

The report included a list of important gaps in the data that will have to be filled if thermal conditions and the IIR Maverick's performance are to be predicted accurately (app. II, item 21).

Figure 30
Threats to Test Quality: The Thermal Character Objective



Threats to test quality

In figure 30, we list the major threats to the quality of the test results on thermal characteristics. There were threats at all five steps from defining the objectives through reporting the results.

Omitted issues and unrealistic test conditions

As we noted above, no friendly ground forces were simulated in the test and, thus, there was no simulation or acknowledgment of the problem of distinguishing friend from foe. The only tank that the friendly forces had as a target was the U.S. M-60 representing the Soviet T-62 tank. In the absence of snow, mud, or other moisture to cool the tracks of the tank in the test, the M-60 may have appeared to be warmer (provided a stronger thermal signature) than the T-62 would in actual combat. The M-60 is about 3 feet higher than the T-62 (10 feet versus 7 feet), and the performance of the IIR Maverick is known to be sensitive to the apparent size of its target. Thus, the omission of friendly ground forces and the use of the M-60 to simulate the T-62 meant that the test did not provide information on how the thermal signatures or the height of the Soviet tank (one that might reasonably be used in a conflict in central Europe) affect the ability of U.S. Air Force pilots to use the IIR Maverick missile against Soviet tanks.

Beyond this, the testing range at Ft. Polk could not provide thermal data representative of the proposed European battlefield. The average soil temperature during the test was 47 degrees Fahrenheit, but the average soil temperature in Europe is well below that. Similarly, only southern pine grows on the test area, while central Europe has a wide variety of forested areas. Consequently, what was learned about the thermal characteristics of the Ft. Polk "battlefield" does not necessarily apply to a European battlefield.

Test changes

The test plan stated that thermal measurements of attacked targets would be recorded for every pass of every mission. This was not done. Instead, infrared photos of typical targets were taken before and after each mission. These photos were used to determine the thermal signature of the right front aspect of a tank when viewed at a
altitude, and this signature was taken as the characteristic of all thermal signatures for all passes of each mission.

The quality of the test results was affected in two important ways. First, not all attacks were against the right front aspect of a tank, not all were at the slant range, and not all were from that altitude, so that the photos do not represent the circumstances of the passes. Second, thermal signatures vary over time and

as weather changes, so that the thermal signature of a tank before and after a mission may not indicate its thermal signature during the mission. Using the infrared photos rather than data from actual passes may have resulted in random error and bias in the test results.

The test plan also noted the importance of obtaining data on the thermal signatures of Soviet T-62 tanks and other "threat vehicles" in operation. The plan was to transport foreign armored vehicles to sites where thermal measurements could be taken at different times of the day, on various terrain, and under varying weather conditions, but this was not done and no explanation was given for the change. Following through as planned would have yielded information about the thermal signatures of Soviet vehicles under a variety of likely conditions that would in turn have provided a point of comparison for the test's thermal signatures and the data necessary for addressing this objective.

Analysis problems

Besides the gaps in the test data because thermal characteristics were not measured as planned, data were missing in the analysis of the limited measurements that were taken. Malfunctioning equipment and scheduling difficulties for four missions prevented making thermal measurements for them. Although this number is small, it includes two of the three morning missions that were flown, so that the results that were reported on thermal signatures do not represent what occurred during morning hours--a time of day when, as the ground warms up, the temperature of everything undergoes considerable variation.

Reporting problems

The JTF reported that launch ranges were significantly shorter when the thermal signatures of tanks were poor. In other words, when there was very little contrast between a target and its background, a pilot had to fly closer to the target in order to find it and, thus, increased the vulnerability of the aircraft to enemy air defenses. The JTF did not mention this in its summary statements about the performance of the IIR Maverick. The failure to emphasize the full implications of an analysis of the IIR Maverick's performance under various thermal conditions detracts from the credibility of the overall conclusions about the operation of the IIR Maverick in combat.

Summary of threats to test results for the thermal character objective

The IIR Maverick was not evaluated in this JT&E for its performance in relation to the thermal characteristics of proposed targets, because Soviet tanks were not used and the testing range was too dissimilar to European terrain. The test was conducted on a simulated battlefield whose thermal characteristics can only doubtfully be generalized to a European environment. The thermal

data that were collected were questionable, and the JTF failed to conclude that there may be problems in the IIR Maverick's performance when the contrast between a target and its background is negligible. The System Planning Corporation stated explicitly that there are many important gaps in the thermal data.

SUMMARY OF QUALITY

The joint operational test and evaluation of the IIR Maverick constituted an ambitious effort to address many complex objectives intended to clarify and resolve important uncertainties about the weapon. Understandably, some objectives were addressed more fully than others, and constraints on time, money, and resources may reasonably have limited the test issues and the ability to test under realistic conditions. However, these limitations and their potential effect on the test results were not fully reported. The presence of test conditions that were favorable to the missile's performance probably led to an overestimation of how the IIR Maverick system operates in combat, but this was not acknowledged by the JTF. This is especially true about target area and acquisition ranges and the probabilities of attacking valid targets.

Despite the omission of important issues and the presence of unrealistic test conditions, the data that were collected in this JT&E appear to be complete and reliable, except for the thermal

Figure 31

Summary of the Quality of IIR Maverick Test Results

Given the favorable test conditions

> --four better-than-average pilots flew all missions in mostly excellent weather, in one small test target area, with the same visual and thermal cues on every pass, with no requirement in the simulation to distinguish friend from foe or to respond rapidly to enemy air defense units--

the test results may overestimate combat capability in terms of ranges and the probabilities of

> target-area acquisition and target acquisition.

Despite the favorable test conditions and the probable overestimation of combat capability, it was determined through the test that

> --workload was a problem on all A-10 missions flown on totally cloudy days;
> --visual cues were essential to success;
> --launch ranges, on the average, did not meet the requirement set forth in the system acquisitions report for missions in poor weather;
> --the time of day, the obstruction of trees, and shallow graze angles made for problems in finding valid targets;
> --poor thermal conditions decreased the probability of success;
> --survivability would be a very serious problem given the attack profiles that were flown.

Yet the JTF concluded that

> the operational test data indicate that the IIR Maverick should meet its operational requirements.

data. Although it is difficult to determine exactly how reliable the various measurements were, the data appear to be within reasonable ranges, and the incidence of missing data is not excessive. The greatest negative factor in the quality of the data was the lack of formal validation criteria for categorizing "for-the-record" passes.

However, the analysis and the reporting of the data were generally poor in quality. The joint test force failed to follow its own analysis plan and to develop criteria for judging success. Some details of the data were not analyzed at all, and others were not analyzed properly. The JTF did not fully report the details of the results, especially those that gave indications of potential technical and operational problems. As a consequence, some details are present but obscured in the report and others are absent. Therefore, the summary that stressed the "impressive capabilities" of the missile system was unwarranted. The conclusion that the IIR Maverick "should meet" its operational requirements is further misleading because of the omission of sufficient qualifications about the test's constraints, especially those that led to the favorable test conditions (including the lack of proper countermeasures testing). We have presented another formulation of this summary in figure 31 on the preceding page.

The effort to collect data that would be high in quality was not supported by the presentation of the findings and conclusions, which do not adequately reflect the information that was made available by the test results. Nevertheless, the test results on the IIR Maverick's performance can be usefully interpreted, even though the test conditions were more favorable than those that would pertain in combat.

THE USEFULNESS OF THE TEST RESULTS

The JTF's intended use

The IIR Maverick JT&E arose from the accountability perspective. Although the members of the DSARC II, part of DOD's accountability system for weapons acquisition, were convinced that the technical feasibility of the system had been demonstrated, they were not convinced that the IIR Maverick was operationally feasible. Therefore, they requested the JT&E in order

> "to more fully understand any operational uncertainties or limitations which may exist and to facilitate the evaluation of appropriate operational tactics during the next phase of the [IIR Maverick] program." (II.C.1, p. 1)

The JTF reported that it had resolved all operational uncertainties identified in the DSARC II deliberations on the IIR Maverick missile and that the test had demonstrated the missile's impressive capabilities. Accordingly, the joint test director stated that the test had contributed to a better understanding of the effect of weather on the IIR Maverick, was useful for procurement

decisions, and provided information for the development of operational tactics for the missile.

We found, however, that the IIR Maverick test results are of limited usefulness because their quality is less than satisfactory. For the reasons we have discussed in this chapter, the data may have overestimated the combat capability of this missile system. We do not agree with the JTF's conclusion that the test resolved the operational uncertainties. Because of the incomplete and inaccurate reporting, the test results are susceptible to misuse. The test data are partially useful, particularly those that reveal potential operational difficulties under certain battlefield conditions, and the test did provide valuable lessons on operational tactics. We believe that the most appropriate use of this JT&E could only follow further analysis and evaluation of the data with full recognition of their limitations.

The IIR Maverick JT&E was timely. The test report met the deadlines set by the DSARC II. For the most part, the test objectives matched the original operational concerns that the DSARC II had raised. Consequently, the test was relevant to the concerns of its requestor. In terms of its security classification, the report was classified "confidential" with a separate "unclassified" executive summary, which allowed a wide distribution of the data, the findings, and the conclusions.

Other uses

The Congress, looking beyond the JTF's conclusions to another review of the data, denied funding for the program:

> "the Subcommittee's House Committee on Armed Services comprehensive review of the data collected during all of the IIR Maverick tests was the basis for the recommendation to deny Air Force request to proceed into engineering development of this seeker during fiscal year 1978." (II.C.26, p. 1)

The Air Force appealed, but the Subcommittee Chairman replied:

> "I again personally reviewed the operational test data and spent several hours discussing the tests with pilots and other knowledgeable Air Force representatives. This second review of the data reconfirmed the deficiencies of this seeker in target acquisition, target lock-on and target discrimination." (II.C.26, p. 1)

The Chairman added:

> ". . . I do not believe it is wise to commit to full scale development of a program that could eventually cost over $1.5 billion. . . . I do recognize, however, your desire to maintain the option to deploy the IIR seeker on the Maverick should future tests indicate that the problems have been resolved." (II.C.26, pp. 1-2)

Thus, one of the major uses of the JT&E results was for fulfill ing the responsibility of congressional oversight. The Congress suggested that another operational IIR Maverick test be conducted with the missile's new centroid tracker in European weather conditions. The first JT&E did not put the operational uncertainties t rest. Instead, it appears to have raised more questions.

These unanswered concerns were addressed in another IIR Maverick test in Europe conducted in January and February 1978, yet even this test did not dismiss the IIR Maverick's operational uncertainties in April 1978: "the system that the Air Force tested is not a system that I would recommend building. We have to make improvements to that system . . ." (II.C.25, p. 2288). While not overly impressed with the test results, the Congress released funds for engineering development, having been assured by OSD that future tests would address the uncertainties.

Was the potential use of the Ft. Polk test realized? Today, more than 6 years later, the operational issues about the IIR Maverick system are still unresolved. Another test program, the IR Maverick Follow-On Test and Evaluation, has been planned, in order to address some of the operational uncertainties first raised at the 1976 DSARC II meeting, including the survivability of the aircraft. The test results did not lead to a change in the program requirements. The estimated cost of the program grew from $1.5 billion in 1977 to almost $6 billion in 1983. Even so, production of the missile has been approved.

CHAPTER 5

THE JOINT TACTICAL AIRCRAFT EFFECTIVENESS AND SURVIVABILITY IN CLOSE AIR SUPPORT ANTI-ARMOR OPERATIONS (TASVAL) TEST

From August 8, 1979, to September 27, 1979, a test was conducted jointly by the Army and the Air Force in the Gabilan and Nacimiento valleys of Ft. Hunter Liggett, California, with the Army deploying AH-1S attack helicopters and the Air Force deploying A-10 fixed-wing attack airplanes, first separately and then together, to support a friendly tank company being attacked by an enemy tank battalion that was being supported by an air defense force. The combat environment was complex and intended to simulate what might occur in a conventional war with Warsaw Pact forces in central Europe in the 1980's. The purpose of the test was to provide data on how well tactical air and artillery units could coordinate their attacks on enemy formations, defeat them, and survive. One of the objectives, called the "synergism" objective, was to learn the effect of cooperative close air support from the Air Force A-10 and the Army attack helicopter units. The findings were to assist OSD in making decisions about acquisition and about force structures and combinations.

About the test, the joint test force reported the following:

"The answer to the question regarding aircraft effectiveness factors was less clear [than aircraft attrition factors]; the only apparent finding being that, taking aircraft attrition into account, effectiveness varied little from one trial site to the other.

"Regarding aircraft attrition, the expected attrition in the Gabilan trials was appreciably higher than in the Nacimiento trials for both the AH-1S and the A-10.

accounted
for the AH-1S attrition while was the
prime contributor of A-10 attrition.

"Regarding synergism employing JAAT [joint air attack teams], expected aircraft attrition for both the AH-1S and the A-10 decreased during JAAT trials. At the same time, aircraft effectiveness was complementary. The mean number of expected Red [i.e., enemy] force casualties produced during JAAT trials was approximately the sum of the casualties produced by the AH-1S and the A-10 operating separately in the AHT [attack helicopter team] and A-10 strike package trials." (II.D.6, p. 16)[1]

---

[1]The bibliographic data for quotations in this chapter are in appendix II, section D.

In this chapter, we evaluate the quality of these findings for the effectiveness, attrition, and synergism objectives. The report of the joint test force is our major focus, but we also consider the analysis reported by the Institute for Defense Analyses (IDA) and reports by the Air Force's Studies and Analyses group and the Army's Training and Doctrine Command.

Overall, the test failed to provide results that can be understood in a European environment, because the test area was restricted to the hot and dry California desert. Moreover, the models and measures for aircraft effectiveness and attrition were not properly validated, so that the results for these objectives must be interpreted with extreme care. As for the effect of employing the Army's helicopters and the Air Force's tactical aircraft together, the test data are questionable, because "synergism" was not appropriately addressed. Despite these serious flaws in the test, TASVAL provided useful lessons for conducting future JT&E's.

THE CONTEXT OF THE TASVAL TEST

TASVAL, a joint test and evaluation of tactical aircraft effectiveness and survivability in close air support antiarmor operations, was requested on September 19, 1977, by the Under Secretary of Defense for Research and Engineering in a memorandum issued to the secretaries of the services. The request was based on OSD's concern, from a management perspective, about the relative advantages of attack helicopters and fixed-wing aircraft. He needed information that would reduce decisionmaking uncertainty arising from the fact that

> "The impact on aircraft survivability of integrated enemy air defenses in the numbers expected in Warsaw Pact force structure, tactics, and doctrine is still largely unknown." (II.D.6, p. 1)

Others were concerned from a knowledge perspective. For example, in appropriations hearings in fiscal year 1979 (on March 15, 1978), an Air Force General stated to the Senate what the services' expectations regarding TASVAL were

> "The A-10 is a very important addition to our capability of supporting ground forces. We just concluded a series of tests with the Army to show how the A-10 and the Army helicopter can operate in the same piece of sky. We have found that they mutually support each other and we get more out of the A-10 and more out of the helicopter when we have the other element there.
>
> "Perhaps it should not have been surprising, but it was something we did not expect and we are going to do more of that kind of testing. A program called Taseval [TASVAL]." (II.D.34, p. 2)

The Army was to be the lead service for TASVAL, which was to be conducted in April, May, and June 1978. A preliminary report was to be due in July 1978, the final report by September 30, 1978. Various delays put TASVAL off until May through September 1979 (II.D.24); the JTF published the final report in May 1980. A short chronology of the TASVAL test program is in appendix V, item 1.

THE TEST OBJECTIVES AND DESIGN

TASVAL was designed to assess three strike "packages" or teams: an attack helicopter team, an A-10 team, and a joint air attack team. Figure 32 lists the final test objectives; they were revised to these three after the test began because of time and instrumentation constraints. Figure 33, the TASVAL test matrix, gives the number of trials that were accomplished for each of the teams, and figure 34 shows the major variables that the JTF designated for the test (the two figures are on the next page). The two sites in the Gabilan and Nacimiento valleys at Ft. Hunter Liggett were instrumented with a range-measuring system for collecting data during the test trials. In appendix V, item 2, we give a description of the test program in greater detail, and in items 3-6 we give details of the training program, the composition of forces, and the methods that were used to collect, assess, and validate the test data.

In the test scenario, the friendly aircraft supported friendly ground forces, which defended against enemy ground forces that

Figure 32

The TASVAL Final Test Objectives

Objective	To evaluate	Pages
Effectiveness	the conditions and factors that most significantly affect the effectiveness of the A-10 (using electro-optical Maverick and GAU-8) and AH-1S (using extended-range TOW), in combination, relative to their separate employment, for destroying armored vehicles in antiarmor close air support.	80-88
Attrition	the conditions and factors that most significantly affect A-10 and AH-1S attrition, in combination, relative to their separate employment, during antiarmor close air support and the types and combinations of defense weapons (as simulated in the test) that appear to extract the highest rates of attrition for the A-10 and AH-1S.	88-94
Synergism	the "synergistic" effect of using AH-1S, A-10, and joint air attack team tactics in antiarmor close air support.	94-98

Figure 33
Design Matrix for the TASVAL JT&E

Team	No. of record trials[a]			No. of valid record trials		
	Gabilan	Nacimiento	Total	Gabilan	Nacimiento	Total
Attack helicopter	11	8	19	7	7	14
A-10	11	8	19	8[b]	7	15[b]
Joint air attack	11	8	19	8	8	16
TOTAL	33	24	57	23	22	45

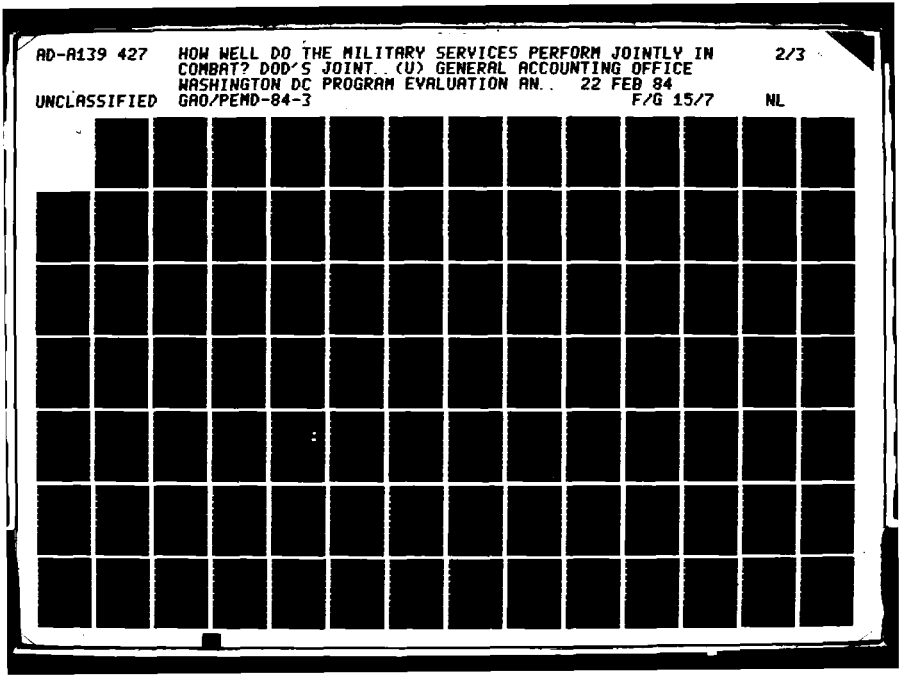
[a]Excludes 20 aborted trials and the one low air defense unit trial.
[b]Two additional A-10 trials were "partially valid."

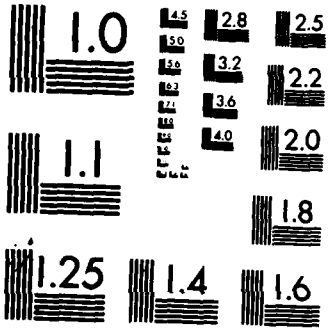
were conducting a breakthrough attack. Figure 35 summarizes the process of providing close air support with the attack helicopters, the A-10's, and the joint air attack team. TASVAL used the kind of two-paired interactive fighting in this process that we described in figure 20 (in chapter 4) for the IIR Maverick. It posed equally difficult problems for simulation in testing and had the same potential for negative interactive fighting (that is, attacking one's own forces).

Figure 34
The Major Variables Considered in the TASVAL JT&E

Independent variable	Dependent variable
The attack helicopter team of 3 OH-58 scout helicopters and 5 AH-1S attack helicopters	Aircraft survivability, measured by the number of aircraft "killed"
The A-10 team of 4 A-10's, one OH-58 helicopter (with forward air controller), and one O-2 helicopter (with rear forward air controller or airborne forward attack coordinator)	The effectiveness of enemy weapons, measured by the number of friendly aircraft engaged or "killed"
The joint air attack team, made up of the attack helicopter team and the A-10 team	Aircraft effectiveness, measured by the number of ground vehicles "killed"

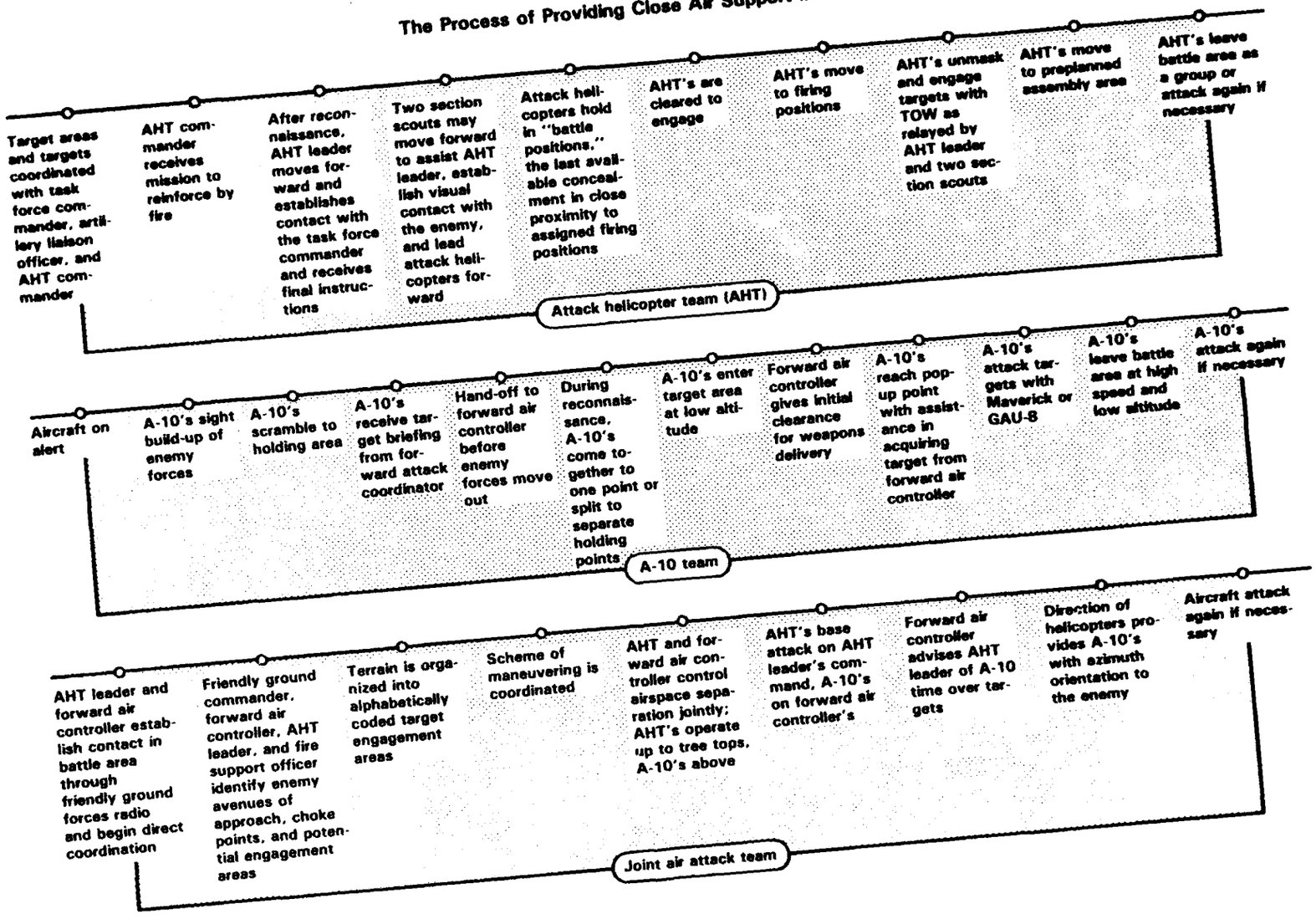
Controlled variable	Uncontrolled variable
Trial starting time	Meteorological conditions
Duration of reconnaissance	Variations in terrain and vegetation
Duration of each trial	Psychological and physiological differences in the players
Rate of enemy advance	Background and experience of the players
Trial site	
Minimum number and type of players required to start a trial	


AD-A139 427 HOW WELL DO THE MILITARY SERVICES PERFORM JOINTLY IN COMBAT? DOD'S JOINT.. (U) GENERAL ACCOUNTING OFFICE WASHINGTON DC PROGRAM EVALUATION AN.. 22 FEB 84
2/3
UNCLASSIFIED GAO/PEMD-84-3 F/G 15/7 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Figure 35
The Process of Providing Close Air Support in TASVAL



THE QUALITY OF THE TEST RESULTS

In the three sections under this heading, we examine each of the final test objectives (listed in figure 32) in terms of how the omission of issues, unrealistic test conditions, test changes, and problems in analysis or reporting affected the quality of the test results. All the quotations of the JTF's objectives and conclusions that we display at the opening of each section are from the report of the TASVAL JT&E issued by the TASVAL joint test force. (The objectives are all on page 3 and the conclusions are all on page 16 of the JTF's report; see document 6, section D, in appendix II of our report.)

Elaboration of test objective and reported results

The JTF presented its results on how well fixed-wing and rotary-wing aircraft can destroy armored vehicles in terms of the mean number of air-to-ground engagements against enemy vehicles per trial and the mean number of enemy vehicles it was expected the aircraft could "kill" per trial. Regarding air-to-ground engagements, the JTF observed that the mean numbers for the two test sites differed only slightly for the attack helicopters and the joint teams but that the A-10 engaged 40 percent fewer enemy vehicles per trial in the Nacimiento Valley than in the Gabilan Valley (app. V, item 7). According to the JTF, friendly aircraft selected enemy armored units as targets more often than they selected enemy air defense units. Approximately 20 percent of the air-to-ground engagements in Gabilan were against enemy air defense units, but in Nacimiento the figure was less than 10 percent. The JTF also noted that the A-10 pilots used the electro-optical Maverick missile more often than the GAU-8 gun and used the missile more often in Nacimiento than in Gabilan. (The electro-optical Maverick is not the same as the imaging infrared Maverick, which was not used in TASVAL.)

Regarding enemy force casualties to be expected from air-to-ground engagements, the JTF observed that the number was about the same in both valleys and for all strike packages (app. V, item 8).

EFFECTIVENESS OBJECTIVE

JTF objective

"Evaluate those conditions and factors which impact most significantly on the effectiveness of the A-10 (using [electro-optical] Maverick and GAU-8) and AH-1S (using Extended-Range TOW) (in combination relative to each employed separately) for destroying armored vehicles in the [close air support]/antiarmor scenarios of the test environment."

JTF conclusions

"The answer to the question regarding aircraft effectiveness factors was less clear; the only apparent finding being that, taking aircraft attrition into account, effectiveness varied little from one trial site to the other."

Regarding which factors affect aircraft effectiveness, the JTF reported that it could not find any.

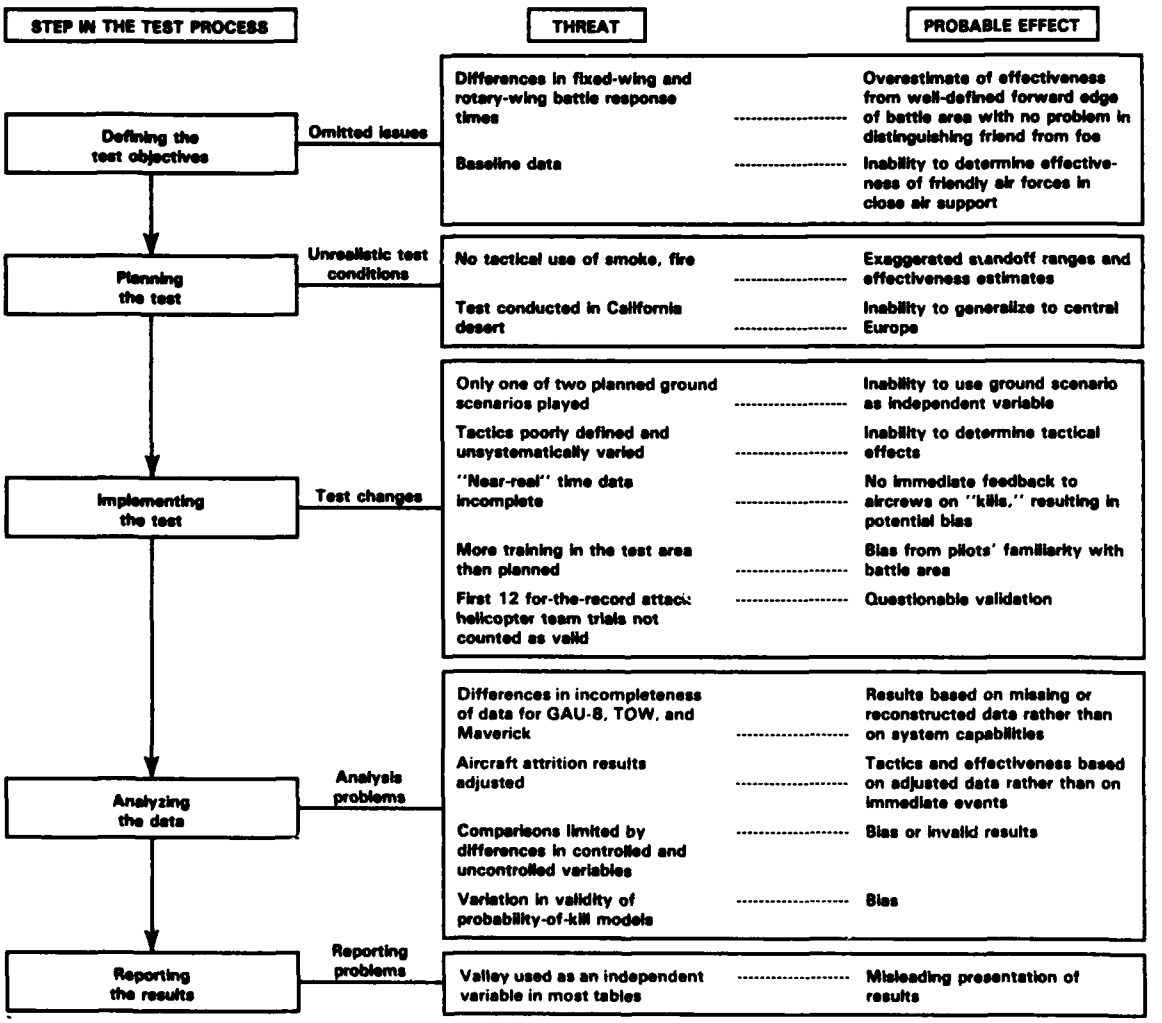
Threats to test quality

Figure 36 summarizes the major threats to the quality of the test results for the effectiveness objective. There were threats at all five steps of the test process from defining the objectives through reporting the results.

Omitted issues

An important omission from the test objectives was the difference in time that it takes fixed-wing and rotary-wing aircraft

Figure 36
Threats to Test Quality: The Effectiveness Objective



to respond in battle to a call for air support. It was assumed that the helicopters and the A-10's would arrive at the battlefield in sufficient time and that they would be early enough to identify and engage the enemy ground vehicles before distinguishing friend from foe became a problem. The problem would become likely if a time delay permitted the attacking enemy force to outnumber the friendly ground forces, advance, and engage them in combat before the arrival of the friendly air forces. An earlier JT&E had already demonstrated differences in the time it takes helicopter and fixed-wing aircrews to respond to calls for close air support.

The 1974-75 Close Air Support Command and Control JT&E showed that the shortest average response time between a request for support reaching a command post and the engagement of the first enemy target is minutes for attack helicopters but minutes for fixed-wing aircraft. If enemy forces in TASVAL had this difference in time to advance without the full threat of the friendly air forces, the opposing ground forces would be mixed, making it difficult for the pilots to tell friendly from enemy forces. Nevertheless, we found no evidence of concern about these command, control, and communication problems in the definition of TASVAL's objectives. Even if it had been thought that incorporating them into TASVAL would have further complicated an already complex test, the 1975 test results could have been used for starting each trial with a random time delay more representative of what might happen in an attack.

Furthermore, the test's objectives omitted critical baseline data. The contribution of friendly air forces in supporting friendly ground forces is not measurable without first knowing how effective the friendly ground forces are in the absence of all friendly air support. Without these baseline data, no comparative analysis could be made of the effectiveness of any air support, but the data were not collected. Without such analysis, it cannot be known whether, for example, the friendly ground forces could have engaged and killed the same number of the enemy with no help from friendly air forces as they killed with help.

Unrealistic test conditions

According to the JTF, no tactical use of smoke or fire could be used in the test because they would have threatened the quality of measurements from the laser pairings that defined an engagement between opposing forces--a firer and a target. IDA's test design specifically stated that "The difficulty in replicating smoke conditions and the possible loss or delays of otherwise good trials is judged to outweigh any realism advantages that might be gained" (II.D.25, pp. 43-44). However, the IIR Maverick JT&E, completed 10 months before IDA's publication of the TASVAL test design, had demonstrated the practicability of employing smoke and fire on the battlefield while simulating close air support. Furthermore, the JTF did not consider any means other than lasers for determining pairings, even though the idea of a conventional

battle without smoke, dust, and fire cannot be considered very realistic. A battlefield with clear visibility could particularly favor optical systems such as the Maverick and TOW missiles that were used in TASVAL, giving pilots considerable opportunity to acquire targets at relatively long standoff ranges. The test's results on friendly air effectiveness may be exaggerated because the targets were easier to acquire in TASVAL than they would be in combat.

The test's location, too, was especially threatening to the quality of the results. IDA's test design stated that Ft. Hunter Liggett's clear skies, its undulant terrain, its high mountains, and its scrub oak trees scattered on bare yellow ground had specific disadvantages, but the chief one is that these are not features that are typical of central Europe. The test designers did recognize that the results of flying over desert cannot be generalized to what it is like to fly over central Europe. However, IDA recommended that Ft. Hunter Liggett be used, in the absence of the ability to obtain a European site, because it had better airspace and better opportunity for electronic countermeasures than other possible sites and was already equipped with some of the instrumentation systems.

Despite the implications of using an unrealistic test site and the impossibility of meeting the test purpose--both known while the test was being designed--the purpose for TASVAL was not changed. It still purported to simulate a heavily defended central European environment. Furthermore, even though one of the reasons for using Ft. Hunter Liggett was its utility for electronic countermeasures, these were deleted from the final formulation of objectives.

Test changes

TASVAL was originally designed to have two independent variables: the type of strike package, or air force team, and the type of ground scenario. Two ground scenarios, a friendly hasty defense and a friendly offense, were planned with the expectation that the two would differ considerably, but only the friendly hasty defense was used. Therefore, what might happen when a friendly force is surprised by an attacking enemy force could not be compared with what might happen when a friendly force has time to prepare a detailed plan of attack against an enemy's defensive operation. The JTF gave time and instrumentation constraints as general reasons for narrowing the scope of TASVAL. Cutting the test in half diminished the usefulness of its results.

Another threat to its quality was the lack of specificity about what aircrew tactics were to be used in the test trials. According to the design, tactics were to be defined and used for a prescribed number of trials so that aircraft losses and the ground targets they killed could be evaluated in terms of both ordnance and tactics. However, tactics were not explicitly defined for each test pass, nor were any plans made to vary them systematically.

Thus, how the various tactics that were used affected the teams' effectiveness cannot be determined.

This is not to say that the tactics could not have been defined. Joint tactics had been developed through another operational test, the Joint Attack Weapons System Tactics Development and Evaluation, conducted in September 1977. However, not only were the joint air attack team's tactics in TASVAL explicitly not defined for each pass; they were also not even coordinated between the attack helicopters and the A-10's. That is, the joint trials that occurred in the test did not represent the tactics that military documentation specifies for joint air attack missions. The JTF did not explain why tactics were not defined or varied according to the original plan, and the quality of the test results is questionable.

To make TASVAL as realistic as possible, the original plan called for all players to be assessed in "near-real" time--an assessment process in which casualties from simulated firings are known shortly after the event and "dead" players cease their participation in the combat or are removed while the test trial is still going on (see appendix V, item 3, on "near-real" time casualty assessment). However, the limitations of the computer prevented the accurate measurement of aircraft performance, which could not then be used in the models to determine whether aircraft were "killed" while a trial was still in progress. Moreover, only the ground players were notified in "near-real" time that they had been "killed"; they were marked immediately by purple smoke and held their positions. The pilots were given no information about kills during trials, and information about firings from enemy air defense units was given to them the next day. This meant that the ground players could learn right away what kept them from getting killed and what did not, while the air players could not learn what worked and what did not because they never knew when they were killed and could learn about when they had been fired upon only a day after it had happened. The test results may be biased in favor of the ground forces, who were able to modify their tactics as the test progressed, or they may be biased in favor of the air forces, if the ground forces wasted fire on aircraft that were already dead without anyone knowing it.

The test results are probably biased in several other ways. First, the participants were involved in an extensive exploratory phase at the test site during their training for the test. Instrumentation problems increased the length of the exploratory phase. It is likely that the pilots were very familiar with the target area. Second, while the services looked on the exploratory period as a chance to revise tactics and techniques in order to improve their performance in the test, the JTF did not report what changes in tactics, if any, were made. Third, when the exploratory phase was to end and when test trials were to begin were not clearly defined. The joint test director decided to begin recording attack helicopter trials in July, but the first 12 trials that were conducted were omitted from the for-the-record group. After

debate about whether they should be included in TASVAL's data base, they were dropped, for reasons of "instrumentation and operational problems," according to the JTF's report. Given the lengthy exploratory phase, in which the attack helicopter team participated in 24 of the 29 pretest trials, it is difficult to understand without further explanation why problems arose after the first 12 record trials that had not been noticed before. The exclusion of these trials without any detailed justification is disconcerting, and the bias that their omission causes cannot be determined, because their results were not published or analyzed.

Analysis problems

In order to determine the effectiveness of the attack aircraft, the analysts had to transform events that took place during test trials to "probabilities of kill." To make that transformation, they first classified firings as valid or invalid, paired or unpaired, and assessed or unassessed (app. V, item 9). A firing was defined as any pull or squeeze of the trigger (gun bursts for the GAU-8, launches for the Maverick and TOW missiles) that was recorded during a trial. To be recorded as valid, a firing had to meet all five of these criteria: (1) the target, if known, was not friendly to the firer (that is, fratricides were not accounted for as valid firings), (2) the firer was alive in real time, (3) the firer had ammunition, (4) the firing was not the result of an instrumentation error, and (5) the firer was following the proper procedures and doctrines.

A pairing was defined as any firing against a target specifically identified by laser, computer algorithm, videotape, or photograph. There was no pairing when only the target type, not a specific target, could be identified. Both paired and unpaired firings that were valid according to the five criteria above were assessed when the outcome of the event was known. Thus, valid paired firings were assessed when there was sufficient information about the firer and the target to assign a numerical value to the event, and valid unpaired firings were assessed when there was ample evidence that no target existed, the firing was therefore a "miss," and the probability of kill was zero.

To be valid, a firing could not be a fratricide. In other words, data on shooting at one's own troops were collected in TASVAL but treated as invalid and not used in the analysis. In combat, it can be difficult to distinguish friendly from enemy forces; when it is, effectiveness is likely to diminish. The analysis may have overestimated effectiveness in battle.

The incidence of paired firing that was reported indicates that instrumentation problems or other factors prevented a systematic assessment of firings for all weapon systems used in TASVAL (app. V, item 10). For example, the A-10 with GAU-8 gun had the smallest number of firings, the lowest overall pairing rate, and the highest pairing rate when identifying the target required some manual adjustment to the instruments.

Instrumentation problems cannot be avoided during a test, but they could have been controlled for by analyzing what differentiated assessed firings based on primary instrumentation from assessed firings based on manual adjustments. This was not done. Therefore, the results that were reported for the GAU-8 gun may be biased, and comparisons of the effectiveness of the attack helicopter team with the A-10 team may be inappropriate. We cannot tell whether the probability of kill estimated for attack helicopters from trials in Gabilan is biased in relation to the A-10's estimated probability of kill in Gabilan (app. V, item 8) because of an invalid measurement of A-10 GAU-8 performance. Nor can we tell whether the A-10's low overall kill rate with the GAU-8, in comparison to the Maverick, is valid.

Since the JTF made no adjustment in its analysis for unassessed firings, it seems to have assumed, without making the assumption explicit, that there was a zero probability of kill for all unassessed firings. In looking for justifications for this assumption, we found that the analysis could have been performed in other ways. One is that of the IDA analysts, who adjusted the data for variations in the completeness of the data base. For example, they assumed that the unpaired and unassessed firings had the same distribution of values for probability of kill as the paired and assessed firings. Neither the JTF nor IDA accounted for the possibility that firings were not assessed because they were unique or idiosyncratic. Furthermore, IDA adjusted the data for 5 percent of the unassessed TOW firings and 4 percent of the unassessed Maverick firings but 24 percent of the unassessed GAU-8 firings, so that the probabilities of kill for the GAU-8 may be more indicative of the adjusted data than of what happened in the test. IDA's procedure might be justified if the assessed firings were not significantly different from the unassessed firings, but this has not been determined.

The final JTF test report stated that the type of strike package or aircraft team would be the only independent variable, but the type of valley was also used in the JTF analysis. Since differences in the way factors related to these variables were not properly controlled for, the results on effectiveness for these independent variables are not really comparable. The attack helicopter, A-10, and joint air attack teams differed in a number of ways, including the following:

AHT	A-10	JAAT	
32.20	35.26	37.20	Mean trial time (hours and minutes)
21.4	18.6	26.2	Mean number of friendly players per trial
76.7	76.9	75.7	Mean number of enemy players per trial
36.0	47.0	63.0	Percent trials before noon
64.0	53.0	37.0	Percent trials after noon
29.0	40.0	50.0	Percent trials with heavy dust
29.0	13.0	44.0	Percent trials with fewer than 3 SA-8 enemy air defense units
14.0	0	37.0	Percent trials with hasty defense in more than one fourth of the trial

For example, it is possible that differences in the effectiveness of the attack helicopter and A-10 teams stemmed from the fact that 40 percent of the A-10 missions but only 29 percent of the attack helicopter missions were conducted in heavy dust. Or it is possible that the helicopter and A-10 teams differed in effectiveness from the joint air attack team because of the time of day when the missions were conducted, while differences in the effectiveness of the attack helicopters themselves may be related to the fact that 57 percent of their missions in Gabilan were in the morning but only 17 percent of their missions in Nacimiento were in the morning. Other differences in the two valleys were as follows:

Gabilan		Nacimiento		
No.	%	No.	%	
4	57	1	14	Attack helicopters at morning
3	43	6	86	Attack helicopters at evening
3	38	4	57	A-10's before noon
5	62	3	43	A-10's after noon
4	67	6	75	Joint air attack team at morning
2	33	2	25	Joint air attack team at evening
12	52	6	27	Heavy dust
4	17	11	50	Tanks and BMP's told not to fire at A-10's
4	38	5	33	Trials after real A-10 crash
5	22	3	14	Hasty defense in one fourth of the trials
12	52	0	0	Defense suppression tactics

All the differences in the mean number of casualties to be expected among enemy forces from the A-10's may derive from the tanks and BMP's having been told not to fire at A-10's during 17 percent of the trials in Gabilan but 50 percent in Nacimiento. (A BMP is a Soviet infantry combat vehicle.)

Some of these factors could have been controlled for during the test's implementation. For example, all three teams could have flown the same number of morning and evening missions in both valleys. Test constraints may have precluded the control of other factors, but then their potential effect should have been analyzed before differences in the trials were attributed either to the type of strike force or to a valley. Without such analysis, it is not possible to state conclusively which independent variable led to the differences in effectiveness, and the test's results are questionable.

One final point about the analysis: major differences in the models that were used to estimate probabilities of kill from data on air-to-ground weapon systems may have biased the results (app. V, item 11). For example, the models for the GAU-8 gun and the Maverick missile assumed stationary targets, but the model for the TOW missile did not. Similarly, the aspect angle was fixed for all targets (except one) in the TOW model but variable for the targets in the GAU-8 and Maverick models. Such differences may be unavoidable, but there was no discussion of how the biases they may have created were analyzed and accounted for.

Reporting problems

As we noted in the section on analysis, the test results on the effectiveness of the three strike teams were reported separately by valley. Without commentary, this presentation may lead the reader to draw unwarranted conclusions about the differences within each valley. The conclusions would be unwarranted if other, uncontrolled factors, such as the differences in numbers of morning and evening missions we discussed above, could also account for some of the differences.

Summary of threats to test results for the effectiveness objective

The results on effectiveness reported by the JTF and the comparisons it made among the three strike teams and between the two valleys are highly questionable. First, since no baseline data were collected, the contribution that friendly air forces made to battle effectiveness, beyond what the friendly ground forces could do alone, was not determined. Second, the results may be biased, possibly overestimating combat effectiveness, because friendly air forces were not required to sort out friend from foe, the lengthy exploratory phase gave the participants considerable familiarity with the test conditions before the test began, ground forces were given immediate feedback on casualties but aircrews were not, certain features of the testing range gave an advantage to friendly air forces that they would not have in Europe, data for friendly ground forces engaged or killed by friendly air forces were considered invalid for analysis, and the first 12 test trials for the attack helicopter team were dropped from the record without explanation. Third, the results may be biased in unknown directions because of missing data, problems with the instrumentation, failure to control for various factors in the test design or the analysis, and differences in the models that were used to estimate the probabilities of kill.

Elaboration of test objective and reported results

The attrition objective addressed the problem of the Army's attack helicopter and the Air Force's A-10 surviving in close air support. One aim of the objective was to identify the enemy defenses that appear to be the most effective against these aircraft. The JTF presented the results as the mean number of enemy ground-to-air engagements against the attack helicopters and A-10's, the mean expected attrition of these aircraft per trial, and the percentage contribution to their attrition from specific enemy weapon systems (app. V, items 12 and 13).

According to the JTF's report, there were more valid enemy ground-to-air engagements against friendly air forces per trial in Gabilan Valley than in Nacimiento, and the mean expected attrition was greater in Gabilan than in Nacimiento, although it differed

ATTRITION OBJECTIVE

JTF objective

"Evaluate those conditions and factors which impact most significantly on A-10 and AH-1S attrition aspects (in combination relative to each employed separately) during the [close air support]/anti-armor scenarios of the test environment. Evaluate which type of threat defensive weapons or combinations of weapons (as surrogated in the test) appear to extract the highest rates of attrition of A-10's and AH-1S's in the test scenarios and environment."

JTF conclusions

"Regarding aircraft attrition, the expected attrition in the Gabilan trials was appreciably higher than in the Nacimiento trials for both the AH-1S and the A-10.
accounted for
the AH-1S attrition while was the prime contributor of A-10 attrition."

for the attack helicoper, A-10, and joint air attack teams. Attrition for the scout helicopter was similar in the two valleys. Exactly what factors accounted for these differences was not made clear.

In reporting the effectiveness of enemy weapon systems against the attack helicopter and the A-10, the JTF presented the as making by far the largest impact against the A-10. The had some impact and the had almost none. As for the attack helicopter, attrition was greatest from

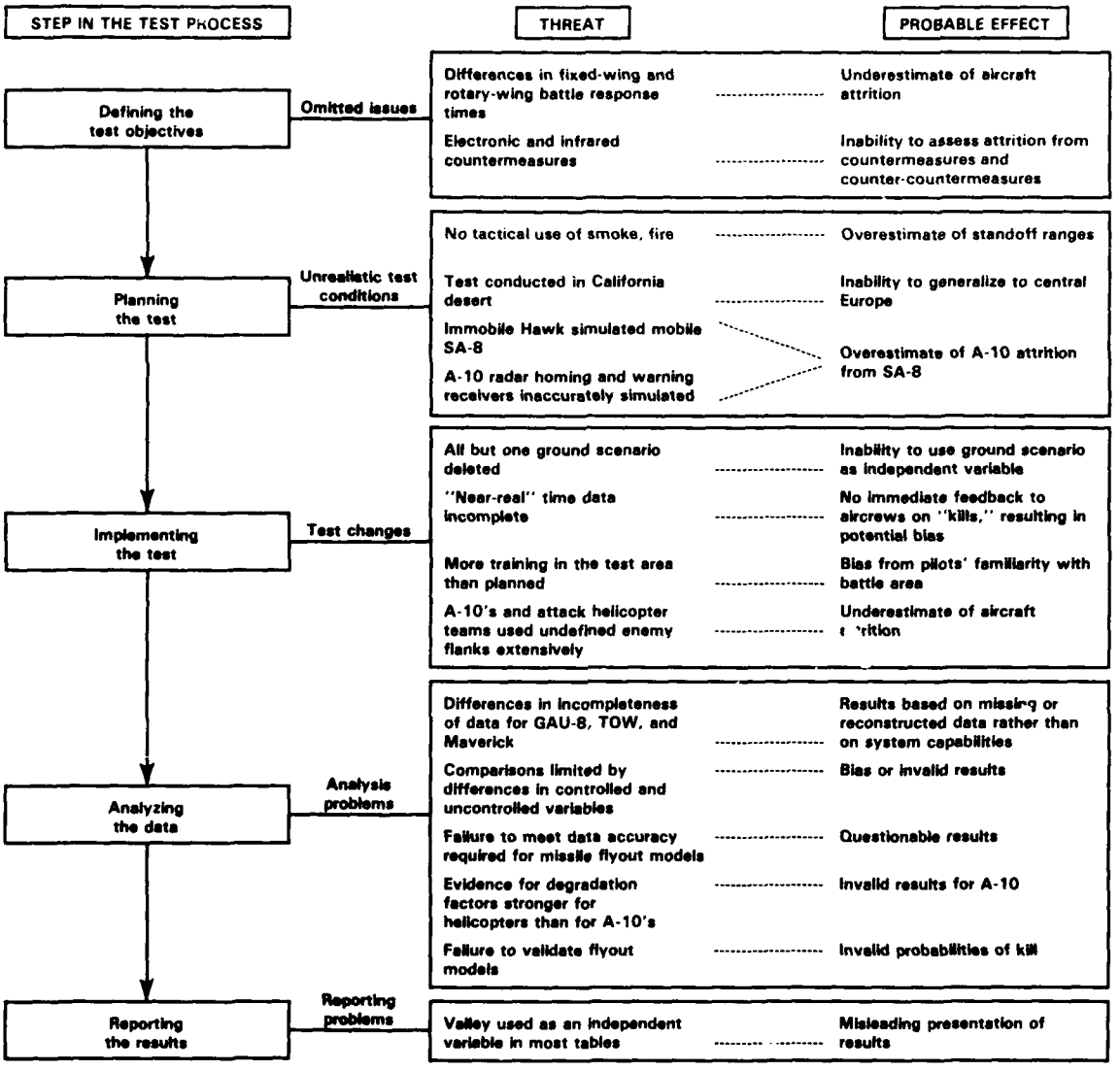
Threats to test quality

Figure 37 (on the next page) shows the major threats to the quality of the test results regarding aircraft attrition. There were threats at all five steps of the test process from defining the objectives through reporting the results.

Omitted issues

We noted for the effectiveness objective that the time friendly air forces needed to respond to the call for close air support from the friendly ground forces was omitted in the problem definition for TASVAL, even though the Close Air Support Command and Control JT&E had shown that a substantial time delay in friendly air forces' arriving at the scene of the battle from their bases might give the attacking enemy ground forces opportunity to overrun the defending friendly ground forces. This, in turn, could create a target identification problem for the pilots, making it difficult to distinguish friend from foe, and leave the friendly aircraft open to exposure to enemy air defense units for a longer time. The omission of the delay in response time means that the attrition rate for the friendly air forces may have been underestimated in the test.

Figure 37
Threats to Test Quality: The Attrition Objective



The effect of countermeasures on combat was dropped as a test objective. No electronic or infrared countermeasures were used. Thus, the interaction of countermeasures and counter-countermeasures was not tested. If it had been tested, the tactics for friendly air forces and enemy air defense units, and the attrition results, might have been different.

Unrealistic test conditions

The omission of smoke and fire in the definition of test objectives created a highly unrealistic test environment,

favoring the effectiveness of the A-10 with the Maverick and the attack helicopter with TOW, missiles that operate best in daylight and good weather. It also provided maximum standoff ranges from enemy air defense units, increasing the likelihood that the friendly aircraft could stay out of enemy range. And Ft. Hunter Liggett's clear weather and small closed valleys, with the terrain and foliage of mountains bordering the desert, were even less generalizable to a European environment.

Furthermore, the simulator for the SA-8 enemy air defense unit, a missile system on wheels, was immobile. OSD had emphasized the importance of testing with the most realistic and credible air defense possible, but changes in scheduling and program priorities ended in the use of the stationary Hawk to represent the mobile SA-8. The results from the stationary simulator, which cannot fire on the move, may overestimate A-10 attrition, since any movement of the real SA-8 might reduce the opportunity of firing at the friendly air forces. It is also possible that the A-10's attrition rates are underestimates, however, since pilots who found the location of the stationary defense unit might have been able to avoid its lethal range.

In addition, the simulators for the equipment that gives aircraft an indication that they are being followed by radar, and the radar's source, differed greatly for the attack helicopters and the A-10's. The equipment for attack helicopters was supposedly well simulated. The equipment for the A-10 did not simulate current capabilities, since the lateness of the decision to use the Hawk meant that there was not enough time to modify the existing gear. According to the Air Force, this posed a severe limitation because

> "First, the pilots maintained that the RHWR [Radar Homing and Warning Receiver] light was difficult to see in bright sunlight and during hard maneuvering due to its location. Second, no radar search warning was displayed at all. Third, normal A-10 RHWR capabilities such as
>
> were not possible." (II.D.6, p. 4-7)

Consequently, it is difficult to determine whether the contribution of the simulated SA-8 to the greater A-10 attrition, in comparison with the attrition of the attack helicopters, is real or merely the result of the differences in the simulated radar-warning equipment for the two types of aircraft.

Test changes

Many of the test changes that influenced the effectiveness objective also affected the attrition objective. The only scenario was for an enemy attack on a friendly defensive position. The aircrews received only delayed information on engagements with the enemy and none on their own survivability. Since they received information on enemy engagements the day after a test

trial, they could not adjust their tactics as they might with more immediate feedback, and the attrition results may reflect this.

The test change that led to the lengthy exploratory phase at the test site may also be reflected in the attrition rates, given that the pilots had opportunity to study the terrain and the ground forces before the simulated combat. The JTF reported that

> "During the many trials of the exploratory phase players learned the terrain, their tactics, the opposing forces strength and general scheme of maneuver and perfected their standard operating procedures. The learning process was enhanced by reviewing the quick look results of the exploratory trials which indicated laser pairings for players. By the time record trials started, the learning (and proficiency) of most players had reached a peak and stabilized." (II.D.6, p. 4-13)

In addition, the test was originally designed so that the aircraft could attack defensive units from the front, with no need, thus, to take undue advantage of enemy flanks. This was done by controlling the approach corridors and the areas in which aircraft would be free to operate by announcing a series of phase lines for the battlefield. However, the pilots were able to reduce the likelihood of their being fired upon by flying over areas where there were few air defense units and yet conduct a valid trial. Apparently the attack helicopter and A-10 teams both took advantage of the areas, the A-10 more than the helicopter teams. This access to flank areas lacking defense may have biased the attrition results toward either or both of the attack aircraft.

Analysis problems

The varying degrees of incompleteness in the data affected the analysis of the attrition results. For example, the percentage of trial events for which no outcome could be determined was greater for ground-to-air engagements than for air-to-ground engagements. For the attack helicopter team, these percentages were 37 and 9, respectively (app. V, item 14). These percentages indicate that instrumentation problems were greater for the ground-to-air pairings. Furthermore, the rate at which enemy air defense units engaged friendly air forces varied from for the ZSU-234 to for the SA-8 (app. V, item 15). Although smoke and fire were omitted from the battle scenario in order that the data on engagements with lasers would be of good quality, the pairing rates that the instruments detected did not rise as high as for any enemy defense unit other than While the had the highest pairing rate, at percent, it was the only enemy defense unit whose pairing rates were based . Therefore, the JTF's conclusion that the was the most effective system against the may not be valid, and the JTF presented no analysis to assure the test's users that it is.

As with the effectiveness objective, IDA normalized the attrition data and the JTF did not. The JTF assumed that the probability of kill for the unassessed firings was zero and did not analyze the conditions surrounding the missing data. IDA assumed that it was equal to the average probability of kill for the assessed firings.

As with the effectiveness objective, differences in attrition between the attack helicopter and A-10 teams could be attributed to factors that were not controlled for. The differences might stem from differences in the length of the trials, for example, or environmental differences in the two valleys, Gabilan being the more dusty. The JTF attributed attrition differences to the characteristics of the teams and the valleys without analyzing the effects of the several uncontrolled variables.

The ground-to-air models that were used to determine how well aircraft could survive posed yet other analysis problems. The instrumentation was not able to pinpoint aircraft locations with the accuracy that the models required, damaging the validity of the analysis. And the factors known as "degradation factors" that the models used to compensate for the lack of countermeasures-testing were based on better and more applicable evidence for the helicopters than for the A-10's. IDA reported that this makes it impossible to tell whether or not the degradation factors represent actual capabilities. The discrepancy in validity between the two models makes it impossible to compare the reported attrition results for the two types of aircraft.

Finally, much controversy surrounded the modeling of attrition from the antiaircraft gun for the enemy ZSU-23-4 and the ground-to-air missile for the enemy SA-7, SA-8, and SA-9. A model that had been validated in an earlier JT&E was not considered at all for TASVAL. The models that were used were not validated, and the TAC-Zinger SA-8 model was changed halfway through the analysis, so that there are two attrition estimates for each weapon. The probabilities of kill reported for TASVAL are highly questionable.

Reporting problems

The attrition results were reported as misleadingly as the effectiveness results. "Valley" was presented in the tables as an independent or controlled variable, despite the fact that differences in the valleys subjected the trials to variation beyond that contributed by the strike teams.

Summary of threats to test results for the attrition objective

Attrition results from TASVAL probably underestimate what could be expected in a European conflict. Friendly air forces minimized their exposure to enemy air defense units by not simulating the possible time delays for arriving at the battlefield,

which also diminished the problem of sorting out enemy from friendly targets, and by not simulating normal battlefield operations in smoke, fire, and dust, which enabled them to fire at maximum ranges, generally outside the enemy's effective defense range. Measuring ground-to-air action was a greater problem than measuring air-to-ground action, biasing the results. The terrain and foliage of the test site were inappropriate in a simulation of central Europe.

Any comparison of how well attack helicopters and A-10's might survive in a conflict in central Europe is questionable if it is based on the TASVAL results. The enemy defense unit was reported as the most effective against

. All the mathematical models that were used for determining aircraft losses were of questionable quality, and those that were used for the A-10 did not represent countermeasure threats as well as did those that were used for the attack helicopter. More data were missing for the A-10, especially its GAU-8 gun, than were missing for the helicopter, because of instrumentation problems. No analysis that would have accounted for the uncontrolled factors was conducted, so that such things as differences in the valleys at the test site and differences in the time of day at which the trials were run make the estimates for the two types of aircraft incomparable in terms of attrition.

Elaboration of test objective and reported results

"Joint air attack tactics" refers to tactics employed by a team of U.S Army attack helicopters, such as the AH-1S, and U.S.

SYNERGISM OBJECTIVE

JTF objective

"Evaluate the synergistic effects of using the A-10 and AH-1S in concert while employing Joint Air Attack Tactics (JAAT) in the [close air support]/antiarmor scenarios of the test environment."

JTF conclusions

"Regarding synergism employing JAAT, expected aircraft attrition for both the AH-1S and the A-10 decreased during JAAT trials. At the same time, aircraft effectiveness was complementary. The mean number of expected [enemy] force casualties produced during JAAT trials was approximately the sum of the casualties produced by the AH-1S and the A-10 operating separately in the [attack helicopter team] and A-10 strike package trials."

Air Force aircraft, such as the A-10, operating together in the same airspace, locating and attacking tanks and other enemy targets in a close air support mission. Of particular interest in TASVAL was whether the total effect of the attack helicopter and

the fixed-wing aircraft was greater when they operated together or when they operated separately. As we showed in the JT&E's design matrix in figure 33, the joint air attack team conducted 16 valid record trials. The team consisted of four A-10's, five AH-1S attack helicopters, four OH-58 scout helicopters (representing a forward air controller, two scouts, and an attack helicopter leader), and one O-2 helicopter (representing a rear forward air controller).

The JTF reported that during joint air attack team trials, the mean expected aircraft attrition decreased or remained the same for every type of aircraft except the OH-58 scout helicopter. The JTF also reported that the mean number of expected enemy casualties during joint air attack team trials was "approximately" the sum of the casualties from the separate operations of the attack helicopter team and the A-10 team.

The JTF further observed that engaging dead targets was a greater problem in the joint air attack trials than in the separate team trials. The number of TOW and Maverick missile engagements against dead targets was greater for the joint air attack team than for the A-10 and the attack helicopter teams (app. IV, item 16). The JTF went on to report that

> "approximately 19 percent of all Air TOW and Maverick engagements were against dead targets in AHT [attack helicopter team] and A-10 strike package trials, respectively. Approximately 29 percent of all Air TOW and 27 percent of all Maverick engagements were against dead targets in [joint air attack team] strike package trials." (II.D.6, p. 3-13)

Although the JTF reported this as "synergism," it never specifically stated whether any total effect was greater than the sum of the parts.

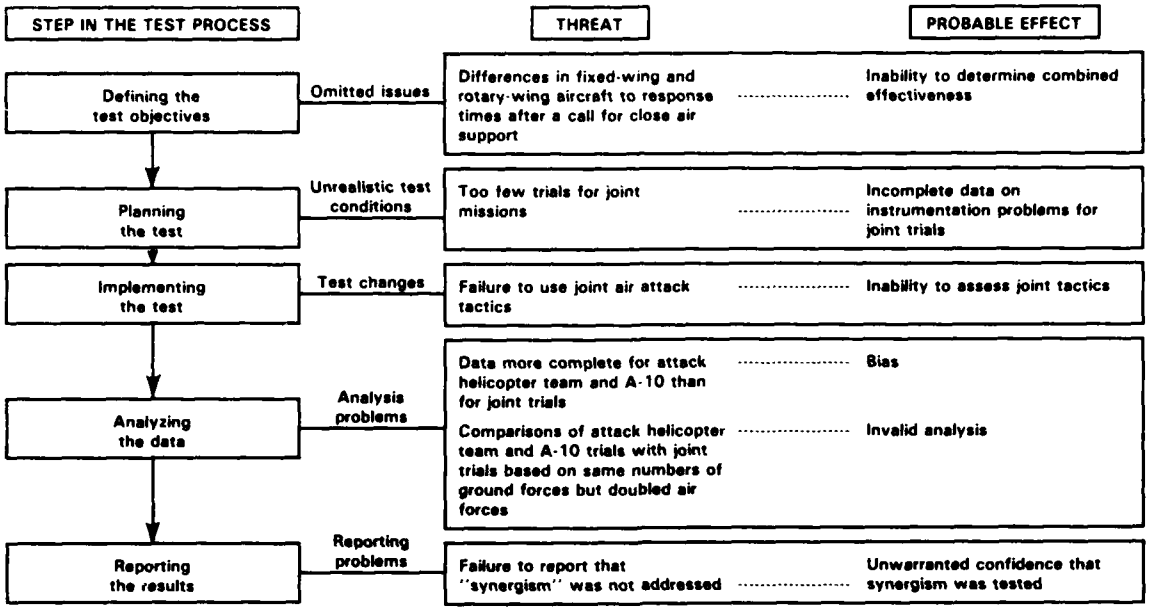
Threats to test quality

Figure 38 on the next page summarizes the major threats to the quality of the test results for the synergism objective. There were threats at all five steps of the test process from defining the objectives through reporting the results.

Omitted issues

The omission of the time needed by friendly air forces to respond to the call for air support is particularly significant for this objective, because it means that obvious questions on synergism could not be addressed. Could the attack helicopter and A-10 both respond to the call for close air support? Could they arrive at the battlefield simultaneously? How much time would they have to operate in the same airspace and how would this affect their operations together?

Figure 38
Threats to Test Quality: The Synergism Objective



The response time from the earlier JT&E on close air support indicated that it takes an average of minutes for the attack helicopter and minutes for the fixed-wing aircraft to respond to a call for close air support. Since the TASVAL trials lasted an average of minutes, allowing for those response times would have meant that the Army and the Air Force could have worked in the same airspace together for minutes. This is much less time for "synergistic" effects to be realized than the minutes that the TASVAL helicopter and A-10's had for operating together. The omission of the response times meant, effectively, that they arrived at the battlefield simultaneously.

Test conditions

The possible difficulty of recording activity during joint air attack team trials, given the greater number of air players compared with trials for the separate teams, was considered but not provided for in the test plan. That is, it was expected that instrumentation failures would be more frequent, but the test plan allowed for virtually the same number of trials for all three teams. Since there were indeed more problems with the joint air attack team trials, 46 percent of their data for ground-to-air engagements was missing, compared with 32 percent for the separate A-10 trials and 37 percent for the separate attack helicopter trials. The number of joint air attack trials was smaller than the number required for a realistic and valid comparison of results from the three teams.

Test changes

The most serious flaw for this objective in the implementation of the test was that no joint tactics were formally defined and none were evaluated, despite the fact that they had been developed by the services. The test documentation and DOD briefings suggest that the only "tactics" that were used were the tactics of mutual noninterference. The airspace was divided into halves, the helicopters flying no higher than the treetops and the A-10's flying above them. They did not get in each other's way, and coordination between the two types of aircraft did not take place.

In November 1977, a joint exercise called Joint Attack Weapons System Tactics Development and Evaluation had been conducted to determine the effects of using the A-10 and the attack helicopter in combined arms operations. Almost two years before TASVAL began, this exercise had demonstrated that current joint air attack tactics are effective when the A-10's change altitude quickly with evasive maneuvers, make effective use of terrain to minimize their exposure to the enemy, and attack simultaneously from several directions and when the attack helicopters use maximum standoff ranges, use the terrain for cover and concealment, and give priority to attacking enemy defenses that threaten friendly air forces, particularly the Of these tactics, aircrews only flew the helicopters at maximum standoff ranges and used the terrain for cover and concealment.

Analysis problems

With the greater number of unresolved air-to-ground engagements in joint air attack trials than in the separate trials, and in the absence of normalization in the JTF analysis, almost half of the data (46 percent) for ground-to-air engagements was not used in the analysis. IDA, using normalized data, substituted average probabilities of kill from the assessed engagements for 46 percent of the joint air attack ground-to-air engagements. The data on aircraft attrition for the joint air attack team are, therefore, questionable.

Despite this, the JTF tried to analyze "synergism" by adding the effectiveness measures in the A-10 trials to those in the attack helicopter trials in order to determine whether the sum represented an effectiveness that was greater or less than that of the joint air attack trials. The comparison was inappropriate. The number of aircraft on the joint air attack team was twice that on the individual teams, while the number of ground forces stayed the same. Only halving the number of friendly aircraft for the joint air attack team or doubling the size of the ground force would have made the comparison meaningful.

The uncontrolled influence of several factors that we discussed for the effectiveness and attrition objectives made comparisons for the synergism objective similarly invalid. For example, more than 60 percent of the joint air attack trials were

conducted in the morning; only 47 percent of the A-10 trials and 36 percent of the attack helicopter trials were morning trials. The JTF analysis did not account for such variations.

Reporting problems

The JTF reported on "synergism" even though it had not been tested in TASVAL. The JTF reported "synergistic" effects during joint air attack trials, in spite of the fact that much data were missing for ground-to-air engagements against the joint air attack team, no specific joint tactics were used, the number of ground forces was the same for all three teams while the number of attacking aircraft for the joint team was twice the number on the separate teams, and the time of day and other variables that could have led to outcome differences were not controlled for. Overall, the results as the JTF reported them are not meaningful for determining whether any effect in any aspect of TASVAL was greater than the sum of the parts.

Summary of threats to test results for the synergism objective

The JTF's comparison of the joint air attack team with the attack helicopter and A-10 teams for purposes of addressing synergism was inappropriate for several reasons. First, it was assumed that the A-10's and the helicopters would arrive at the battlefield at the same time, even though it had been demonstrated before TASVAL that simultaneous arrival may not always be possible in combat. Second, greater instrumentation problems during the joint air attack trials made greater gaps in the data for them than for the A-10 and attack helicopter trials. Third, the joint air attack team used no specifically defined joint tactics, and there was very little, if any, coordination between aircrews of the A-10 team and the attack helicopter team. Fourth, uncontrolled factors that may have affected the results were not analyzed. Fifth, and most important, the joint air attack team had twice as many aircraft as the separate attack helicopter and A-10 teams, while the number of ground forces remained the same for all three teams. The JTF's report on synergism is not appropriate.

SUMMARY OF QUALITY

TASVAL represents a very ambitious effort to address many of the complexities of conventional close air support. Unfortunately, the test failed to address its objectives adequately, even after its basic objectives had been reduced and revised. Figure 39 summarizes our observations about TASVAL's quality. The test did not accomplish what it set out to do, but it did apparently teach many valuable lessons about conducting large-scale force-on-force tests and evaluations. It is not clear, however, whether these lessons will be used to the benefit of future testing.

Although a very detailed and lengthy test planning and pretesting phase took place, certain important issues were omitted

Figure 39

Summary of the Quality of TASVAL Results

The test set out to

reduce uncertainties associated with decisions on weapon systems acquisition and the structure and combination of forces by evaluating the ability of the Army and Air Force to provide close air support in a conventional war in central Europe.

A detailed test planning process took place in which

the issues were identified, the requisite instrumentation was determined, the necessary resources were delineated, and the test parameters were defined.

Unforeseen but predictable circumstances led to the failure to

fully develop essential measuring instruments, assess "near-real" time casualties for all test participants, collect baseline data, train and pretest in the test area for an appropriate length of time, use verisimilitude in choosing the test site. and base outcome measures on validated models.

Consequently,

very little if anything can be said about the ability of the Army and Air Force to provide close air support in a European environment, and

the reported results on operational effectiveness and aircraft survivability are questionable, even for the California test site, because

--issues such as the time aircraft need to arrive on the battle scene were omitted,
--the lack of smoke, fire, and dust made for unrealistic combat conditions,
--the test data were incomplete,
--uncontrolled and unanalyzed factors may have invalidated the comparisons, and
--mathematical models used to estimate effectiveness and attrition were not validated.

Fortunately, this testing experience

taught some important lessons about the process of testing such complex issues.

from the test. The time aircrews need to respond to a request for close air support in battle was not considered, for example. Past testing on this response time had made information available that could have been used in TASVAL, but it was not.

The failure to choose an environment for the test that would represent central Europe means that the test results cannot be applied to a European environment with any predictability, negating its purpose. Other unrealistic test conditions also detracted from the quality of the test results. For example, in order to use laser instruments for engaging targets, normal battlefield smoke, fire, and dust were omitted. Ironically, the laser pairing system was not very successful.

The composition of the three strike forces was the independent variable, but certain features of the test site and other factors that were not controlled for or not analyzed thoroughly threatened the validity of making comparisons among those forces. Differences in the two valleys at the test site were sometimes reported but not always accounted for appropriately in the results. Some of these factors could have been controlled for in the test

design; those that could not should have been addressed in the analysis. They were not.

As for the completion of the test objectives, the way in which "effectiveness" and "attrition" were stated dictated the use of mathematical models to convert test events to "probabilities of kill" and "attrition rates." However, the test instruments failed to provide data that would meet the requirements of the models. Moreover, the models had not been validated and were based on different assumptions. Consequently, the validity of the effectiveness and attrition estimates is highly questionable. Results for the "synergism" objective were reported, despite the test's having revealed nothing about whether the effect of any sum was greater than the parts. Not very much of the TASVAL test data meets our test for quality.

THE USEFULNESS OF THE TEST RESULTS

The JTF's intended use

The information from TASVAL was to help OSD make decisions on weapon systems and force structures and combinations as they might be used in typical close air support missions in a heavily defended area in central Europe. The Secretary of Defense had raised these issues in 1977, requesting that the results be available by September 1978. The TASVAL test trials were not completed until the following September. The joint test force did not publish its report until May 1980. TASVAL was not timely.

Even if TASVAL had been completed on time, however, it could not have usefully served its purpose, because the results do not apply to any European environment. They are restricted in their utility to the hot and dry valleys of the testing range at Ft. Hunter Liggett in California. Therefore, the test results are not relevant to what the requestor asked for.

Even if TASVAL had been completed on time and were relevant to its purpose, it still could not be used to address questions about the acquisition of weapon systems and the combination and structure of forces. Questions about weapon systems acquisition cannot be addressed from TASVAL's data because they allow no valid comparisons of the three strike teams. Questions about the combination and structure of forces cannot be addressed, because force combinations and structures were not systematically varied during TASVAL.

Other uses

The three memorandums we quote in appendix V, item 17, all addressed by DOD officials to the Deputy Director for Defense Test and Evaluation, spell out a number of uses for TASVAL beyond those intended by OSD. The joint task force also compiled and published the "lessons learned" from TASVAL (II.D.14). The uses fall into three categories: tactics, training, and testing. In terms of

tactics, TASVAL reinforced the U.S. Army's priority for using the scout helicopter in security and reconnaissance roles and as an air defense warning and a decoy for other aircraft. For the U.S. Air Force, TASVAL reinforced the importance of the A-10's using terrain for masking and low-altitude tactics. The Air Force noted that its belief that defense suppression is vital was reinforced and that there is positive benefit in using the A-10 and the attack helicopter as a "joint team."

In terms of training, both the Army and Air Force found that TASVAL provided beneficial lessons. In particular, TASVAL helped them recognize the need for improving joint air attack training and tactics.

In terms of testing, both the Army and the Air Force conducted their own analyses of the TASVAL data and learned many valuable lessons about managing large-scale force-on-force tests (II.D.8 and 14-21). The Army reported that what was learned in TASVAL led to improvements in the testing of the AH-64 helicopter and in phase II of the Electronic Warfare During Close Air Support test, among other things. The problems with TASVAL's instrumentation were usefully considered in planning two subsequent JT&E's, and so were the problems with TASVAL's preparation, operations, logistics, and control and the management of its data. In addition, the Army stated that great advances were made in the development of near-real time casualty assessment and removal, although the Air Force reported that the effectiveness of the A-10 attrition rates in TASVAL were highly dependent on the test's scenarios.

CHAPTER 6

THE MULTIPLE AIR-TO-AIR COMBAT (ACEVAL) JT&E

From June to November 1977, in a series of mock air combat trials at Nellis Air Force Base, Nevada, the Navy flew its F-14 and the Air Force flew its F-15 as friendly forces against a common threat, the F-5E simulating an enemy MIG-21J. The JT&E, called Multiple Air-to-Air Combat, or ACEVAL, was meant to represent a fighter sweep mission within visual range, with aircraft encounters ranging from one on one (1v1) to four on four (4v4). The availability of information (called "ground control intercept" or GCI information) telling pilots the relationship between their position and the source of specific enemy threats was varied, and so was the force ratio: sometimes the friendly force outnumbered the enemy, sometimes the enemy force outnumbered the friendly, and sometimes they were even.

ACEVAL was potentially very important. It was thought that a clear picture would emerge of how air combat depends on the number of aircraft that are engaged on each side under various conditions. Such information would be useful to OSD and the services in making decisions regarding fighter aircraft and force structures. It was hoped that ACEVAL would also demonstrate whether its methodology could be applied to other highly instrumented operational tests of multiple aircraft situations in combat. In particular, the question was raised of whether the data derived from such testing could be used to make projections about larger, untested force structures.

ACEVAL was generally successful in demonstrating that a highly instrumented operational test of air-to-air combat can be conducted. The question of projecting ACEVAL data to untested force structures was not really addressed, because known limitations prohibited it. However, several critical aspects of ACEVAL's design, implementation, analysis, and reporting lead us to question the results that the JTF reported, even with its description of the test's constraints and qualification of the results. Relatively few test trials were conducted with the maximum number of aircraft (4v4), and several problems were encountered in those that were, so that it may not be appropriate to compare these test trials with the one-on-one trials. The ground control intercept information that was available to the aircrews was much greater than would be available in combat; as a result, the advantages in having it that the JTF reported may be overstated. The capabilities of the friendly forces may also be overstated, given that test equipment and instrumentation may have favored them. Finally, the F-14 and F-15 may have differed because of the way the test trials were conducted, not because of differences in the aircraft themselves.

For our review of ACEVAL's quality, we focused on the JTF's four-volume final report (app. II.B.12-15) and on its report entitled Air Combat Evaluation Test: Management Lessons Learned

(app. II.B.7).[1] We also used reports issued by the Air Force and the Navy, the Institute for Defense Analyses, independent analysts, and others (all listed in appendix II). Since the JT&E called Air Intercept Missile Evaluation (AIMVAL) preceded ACEVAL and used the same test procedures, we examined the reports on AIMVAL as well (see appendix II.B). Originally, ACEVAL was to be implemented first, in order to test the methodology and to establish an initial data base. However, DOD's need to fulfill its commitment to the Congress to test and evaluate various missile concepts before developing the engineering of a new short-range air-to-air missile dictated that AIMVAL be conducted first. A brief chronology of the ACEVAL test program is in our appendix VI (item 1), along with other technical data on ACEVAL.

In the rest of this chapter, we first discuss the context in which ACEVAL took place and then present a short description of its objectives and design, which are given in more detail in appendix VI (items 2-4). Next, we present our observations about the major threats to the test quality for the test's objectives in terms of the first six steps of the test process. Finally, we summarize our observations about the test's quality, before concluding the chapter with our observations about step 7, the usefulness of the test's results.

THE CONTEXT OF THE ACEVAL TEST

The ACEVAL test program was carried out in response to a general need for information identified by OSD's Program Analysis and Evaluation staff. In April 1974, the DDT&E explained the initial expectation for the ACEVAL test program to the Defense Subcommittee of the U.S. Senate Committee on Appropriations:

> "Estimates of the effectiveness and losses in air-to-air combat to date have been based generally on one-on-one aircraft engagements, extrapolated by mathematical models to evaluate the more frequent multiple-on-multiple, situations. There are real questions as to the validity of such extrapolation. This test's objective is to obtain measured test data on typical multiple-on-multiple combats." (II.B. 25, p. 11)

The DDT&E's request for a study to determine the likelihood of obtaining such data stated why the information was needed:

> "A capability to make assessments of the relative effectiveness of U.S. fighter aircraft against enemy aircraft in air combat is required as a basis for decisions concerning the size and composition of the future fighter force." (II.B.20, p. 1)

---

[1]The bibliographic data for the citations in this chapter are in appendix II, section B.

Realizing that a single test could not appropriately address all issues concerning the outcomes of multiple air combat, the Weapons Systems Evaluation Group and Institute for Defense Analyses proposed narrower objectives for ACEVAL. The test's scope as defined in the feasibility study was limited to a series of air-to-air combat flights that would determine how fighter aircraft losses in encounters between specific aircraft systems within visual range of each other are related to the number of aircraft on each side, given specific initial conditions. The DDT&E accepted this recommendation as addressing the original analytic objectives. It was understood that estimating the size and composition of fighter forces for multiple air combat would require other tests using ACEVAL's procedures.

Since no forthcoming decision prompted the recommendation for the test, its purpose was exclusively to develop information. It falls into the category of tests and evaluations performed from a knowledge perspective. That is, ACEVAL was to derive an empirical data base from an operational test program that could be used later to answer procurement and management questions about the size and composition of U.S. fighter forces. Although no time constraints were imposed, ACEVAL was to be completed during 1976.

THE TEST OBJECTIVES AND DESIGN

The JTF addressed the test objectives listed in figure 40. The design matrix and major variables considered in ACEVAL are given in figures 41 and 42. As these figures show, ACEVAL was made up of two separate experiments or series of flights, each consisting of 360 valid trials. The F-14 in one series and the F-15 in the other flew a fighter sweep mission against and

Figure 40

The ACEVAL Test Objectives

Objective	To provide data on	Pages
Aircraft numbers	how the number of aircraft on each side determines the outcome of air-to-air encounters between specific aircraft systems within visual range.	107-15
Ground control intercept	the effect of pilots' having information about the relationship between their position and the source of specific enemy threats.	115-18
Aircraft type	how outcomes differ by aircraft type.	118-21
Combat elements	how the primary control variables affect combat elements such as aircrews, hardware, and "key" activities such as detecting, identifying, and killing the enemy.	122-23
Test effectiveness	the effect of the test's constraints and the effectiveness of its procedures.	app.VI item 2

Figure 41
Design Matrix for the ACEVAL JT&E[a]

Force ratio friendly to enemy	Encounter size friendly v enemy	Ground control intercept advantage			Total trials
		Friendly	Neutral	Enemy	
1:2 (0.5)	1v2	18	24	18	60
	2v4	9	24	9	42
1:1 (1.0)	1v1	24	24	24	72
	2v2	12	24	12	48
	4v4	6	24	6	36
2:1 (2.0)	2v1	18	24	18	60
	4v2	9	24	9	42
TOTAL		96	168	96	360

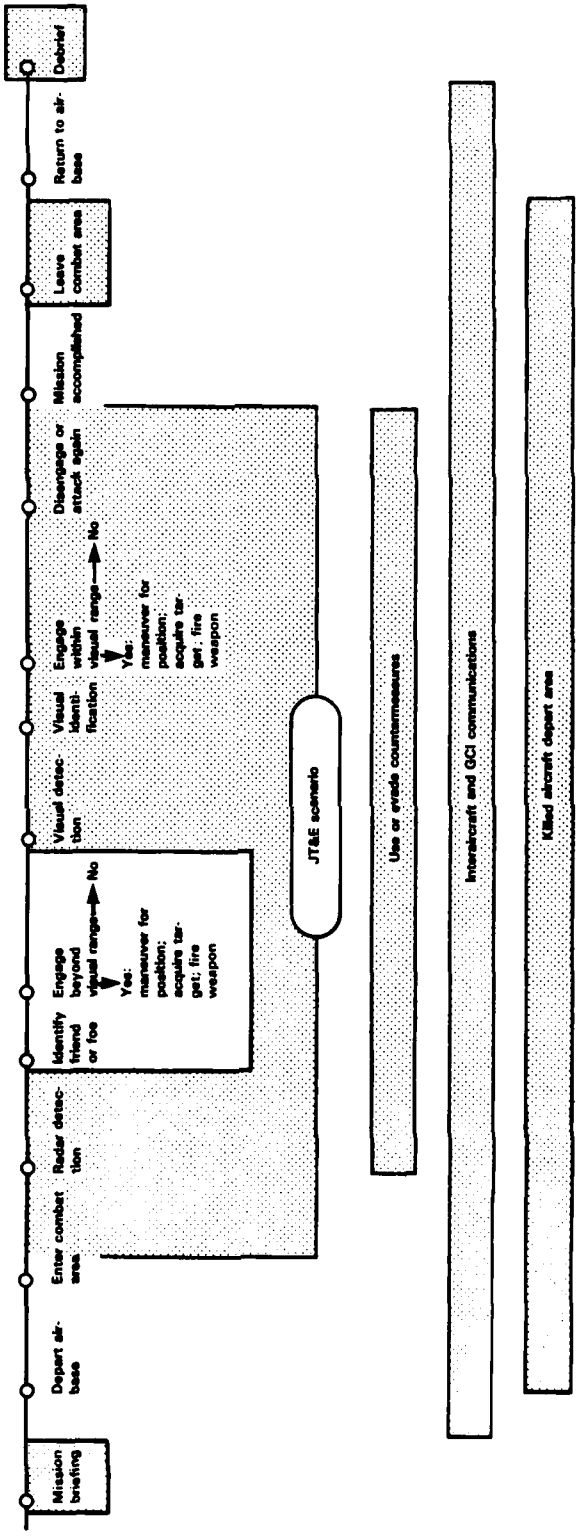
[a]Two concurrent series of trials were flown against the F-5E, one with the F-15 and the other with the F-14, for a total of 720 test trials.

within visual range of an enemy force, F-5E aircraft simulating the Soviet MIG-21J. The number of trials was to vary for each combination of force ratio, encounter size, and GCI. In addition to these 720 trials, 44 air-to-air combat trials included neutral "intruders" with dissimilar aircraft to enforce the problem of identifying targets visually before firing weapons. All trials were flown on the air combat maneuvering instrumentation range northwest of Nellis Air Force Base. Figure 43 on the next page shows the process of air-to-air combat in ACEVAL.

Figure 42
The Major Variables Considered in the ACEVAL JT&E

Independent variable	Dependent variable
	Primary Measures of Effectiveness
Force ratio, or the number of aircraft in one force divided by the number of aircraft in the other force in a trial (1:2, 1:1, 2:1)	Loss rate, or the number of aircraft on one side killed in a trial divided by the total number on that side at the beginning of the trial
Encounter size, or the number of aircraft in a trial in terms of the number of friendly "versus" the number of enemy (1v1, 2v2, 2v1, and so on)	Exchange ratio, or the number of aircraft killed on one side divided by the number killed on the other
Ground control intercept information Neutral, available to both sides Enemy advantage, available only to the enemy Friendly advantage, available only to friendly forces	Other Measures of Effectiveness Percentage of trials in which a force made the first radar or visual detection, visual identification, firing, and kill and the distance between opposing aircraft when these key activities occurred
Aircraft type F-14 F-15	Number of radar or infrared missiles fired and intercepted per kill
	Percentage of opposing aircraft targeted
	Percentage of friendly aircraft firing
	Percentage of kills in time

Figure 43
The Process of Air-to-Air Combat in ACEVAL



THE QUALITY OF THE TEST RESULTS

In the four sections under this heading, we examine each of the test objectives listed in figure 40 in terms of how the omission of issues, unrealistic test conditions, test changes, and problems in analysis or reporting affected the quality of the test results. All the quotations of the JTF's objectives and conclusions that we display at the opening of these sections are from the report of the ACEVAL JT&E issued by the ACEVAL-AIMVAL joint test force. (The objectives are all on page II-1 and the conclusions are all on pages II-2 through II-6 of that report, unless noted otherwise; see document 13, section B, in appendix II of our report.)

Elaboration of test objective and reported results

To accomplish the objective of determining the effect of the number of aircraft on air combat, the test was designed for 720 trials in two sets of 360 (as the design matrix in figure 41 shows). In each set of 360, both flown against the "enemy" F-5E, one with the F-14 and the other with the F-15, the opposing forces were each outnumbered in 102 trials. That is, in 60 of these trials, one aircraft (friendly or enemy) flew against two of the opposition (enemy or friendly) and in 42 of these trials two aircraft flew against four of the opposition--the total of the two sets together was 204 trials. These differences varied the force ratio. In the remaining 156 trials, the encounter size was varied while the numbers of friendly and enemy aircraft were equal, one against one, two against two, or four against four. The numbers of these trials varied, decreasing from a high of 72 trials of 1v1 to the low of 36 trials of 4v4.

The availability of ground control intercept information was also varied. Both forces had it for about 50 percent of the

AIRCRAFT NUMBERS OBJECTIVE

JTF objective

"Determine how the outcome varies with force ratio" and "whether the outcome varies with encounter size for constant force ratio" for each ground control intercept condition.

JTF conclusions

For force ratio, "In the ACEVAL scenario, being outnumbered was the most dominant factor in causing increased loss rates for the outnumbered side. The side with superior numbers considerably reduced its loss rates. . . . The observed adverse force ratio effects were primarily attributed to hardware factors." For encounter size, "For a given force ratio, as the number of aircraft on each side increased, there was a decrease in the [enemy] loss rates while the [friendly] loss rates remained the same (except 4v4 engagements where [friendly] loss rates marginally increased. . . . The observed diminishing returns of a weapon system advantage with increasing numbers (given a constant force ratio) were attributed primarily to human factors."

trials, giving neither side an advantage. The friendly force had it when the enemy force did not for about 25 percent of the trials, and vice versa for the remaining 25 percent.

The outcome of the mock air combat was measured in terms of loss rate (the number of aircraft on a side that were "killed" in a trial divided by the total number of aircraft on that side at the start of the trial) and exchange ratio (the number of aircraft on a side that were killed in a trial divided by the number of killed aircraft belonging to the opposition in that trial). The JTF's results showed that the outcome in exchange ratio was directly proportional to the force ratio for a given friendly flight size (app. VI, item 5). For example, when the friendly force flew two aircraft, the exchange ratio increased from as the force ratio increased from 1:2 (2v4 trials) to 1:1 (2v2 trials) to 2:1 (2v1 trials). The effect of force ratio for trials with two friendly aircraft is illustrated in figure 44, in which it can be seen that the force that was outnumbered had the greater loss rate (see II.B.13, p. VI-11). However, the loss rate for the friendly aircraft was less sensitive to the effect of force ratio than the enemy loss rate. The JTF reported that this difference was highly significant. Given the aircrews' perceptions and indirect measures, the JTF attributed decreases in the effectiveness of friendly forces to the limitations in their weapon systems--namely, the requirement that the AIM-7 missile track a target until it is intercepted and the long fire-to-intercept time.

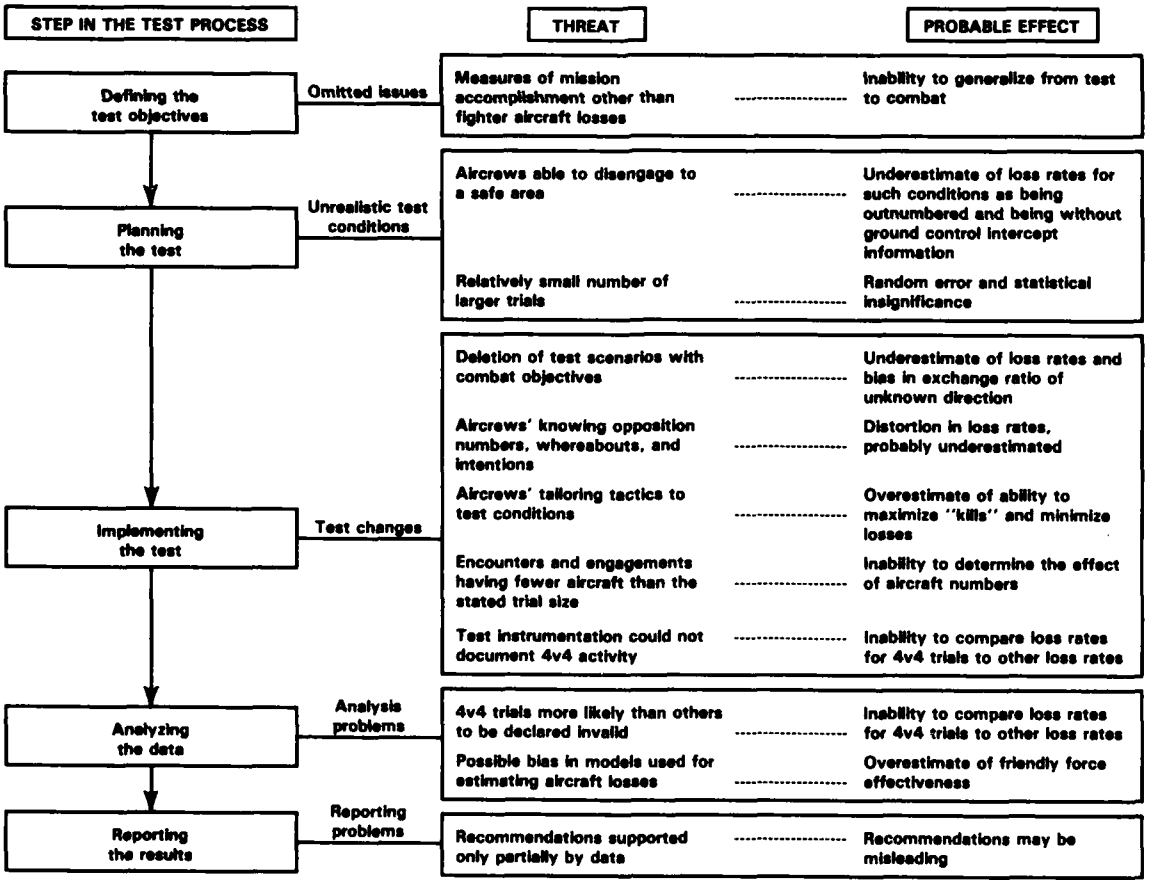
Figure 44
Loss Rates for Force Ratios for Trials with Two Friendly Aircraft (Normalized for GCI)

Regarding encounter size, the JTF reported that the exchange ratio tended to decrease for every force ratio as the number of aircraft increased. For example, for the 1:1 force ratio, the exchange ratio decreased from as the flight size increased from one to two to four (see the 1v1, 2v2, and 4v4 trials in app. VI, item 5). Enemy loss rates were affected the most by increases in encounter size. As the number of aircraft increased, enemy losses rose while friendly losses remained relatively the same, except for the 4v4 encounters, in which friendly losses increased marginally. The JTF attributed the diminishing of the friendly exchange ratio as aircraft numbers rose to human factors such as confusion and loss of coordination, and the attribution was supported by the aircrews as well as by indirect measures for weapon firings.

Threats to test quality

The greatest threats to the quality of the test results for the aircraft numbers objective are summarized in figure 45. There

Figure 45
Threats to Test Quality: The Aircraft Numbers Objective



were threats at all five steps of the test process from defining the test objective through reporting the results.

Omitted issues

In the test's feasibility study, the Weapon System Evaluation Group and the Institute for Defense Analyses (WSEG/IDA) defined the fundamental interest to be the number of fighter aircraft lost and surviving on both sides. It was this measure of effectiveness that was adopted rather than all or any of several other measures of overall mission accomplishment, such as restricting or denying the enemy's air movement or preventing the enemy from disrupting one's operations. All such concepts of air combat success were excluded from ACEVAL. Acknowledging that the basic measure of effectiveness in air combat is the degree to which overall mission objectives are accomplished, WSEG/IDA pointed out that "the ability to attain some of these objectives cannot be measured solely by consideration of the results of air-to-air combat between fighter aircraft" (II.B.26, p. 6). Nevertheless, it omitted overall mission achievement in the definition of test objectives, and the omission affected the quality of the results. Both the JTF and the services reported that the test results reflected the aircrews' gamesmanship, not what they might do in actual air combat. Perceiving that the outcome to be measured was stylized, they used tactics that would win the test "game" rather than achieve the broader purposes of a combat mission.

Unrealistic test conditions

Aircrews in the test were allowed to disengage from the mock combat into a "safe" area near the test range whenever conditions seemed threatening or unfavorable. Aircrews in real combat can never assume that they are moving into undefended or battle-free areas. The JTF and IDA both pointed out that the lack of realism enabled ACEVAL aircrews to reduce their loss rates significantly, simply by disengaging. Loss rates for friendly aircraft that did not disengage were at least twice the rates for those that did; enemy loss rates were five times greater when their aircraft did not disengage (app. VI, item 6). Aircrews disengaged more often when the force ratio and GCI condition were least favorable, and it cannot be known whether their loss rates under these adverse conditions would have been higher if they had not been free to disengage.

The 4v4 trials were qualitatively the most realistic in that they came closer to reflecting the larger encounters, with the concomitant uncertainty and confusion, that are characteristic of combat, but the results from these trials are questionable. There were fewer of these trials than for the smaller encounters, and they had greater problems in test implementation and trial validation. Whereas there were 72 1v1 trials, there were only 36 4v4 trials, and the difference was not consistent with the test matrix that was originally proposed.

WSEG/IDA had proposed originally that the numbers of trials differ but that they be larger as the number of aircraft rose. In the development of the test design, however, this approach was dropped in favor of two assumptions. The first was that when neither side had an advantage regarding ground control intercept information, the number of trials for all encounter sizes should be equal. It was assumed that, in this "neutral" case, advantage would accrue by chance, rather than by control or specification, and that outcomes would vary independently. This would help satisfy both the concern for statistical independence and the logical need to observe what happens.

The second assumption was based on the general reasoning that the number of aircraft exposed to risk (that is, involved in the encounters but lacking ground control intercept information) should be as nearly constant as possible. The number arrived at for each combination of encounter size, GCI advantage, and aircraft type was approximately 24 aircraft to be exposed to risk when only one force had GCI information. However, this meant that the total number of trials for the larger encounters was much less than for the smaller ones. When either side had the advantage, the assumption allowed only 12 4v4 trials to be flown but 48 1v1 trials. The assumption implied that a 4v4 trial is no more complex and can have no other outcomes than a 1v1 trial.

In reality, aircrews fight one way when they know the number of aircraft involved, and when they know it is small, and another way when the number of aircraft is unknown or there are too many to keep track of. As the number of aircraft increases, so do the complexity of the combat and the variety of possible engagements. The number of larger trials was too small for the purpose of comparing more complex encounters with the simpler ones. In its report, the JTF stated that "The assumptions in the test design were specifically used to reduce the required number of higher force mix trials, which were correctly perceived to be the most difficult and costly to obtain" (II.B.13, p. VII-17). Indeed, the larger trials were the most difficult to implement, but the purpose of the test had been expressly to determine what happens in multiple aircraft encounters. Given that combat in the Middle East had recently indicated that much larger encounters, involving 12 aircraft or more, were likely to become the norm, WSEG/IDA's original proposal would probably have provided results of better quality because it matched the purpose more realistically.

Test changes

The test design stated that ACEVAL should include the element of surprise that is found in real air-to-air combat. The aircrews were to know only the starting conditions and the opposition's size and were to have learned only the tactics generally to be expected during air-to-air combat. The design was not adhered to. For one thing, the aircrews began ACEVAL exceptionally knowledgeable about their opponent's abilities and limitations, because

they had participated together for six months in AIMVAL. For another, friendly and enemy forces were stationed at the same air base, and it was extremely difficult, if not impossible, to keep them from sharing information. Further, the aircrews, knowing their starting points and GCI condition, were already aware of where an enemy aircraft might be. By several quick maneuvers at the start of a trial, they were able to determine accurately the enemy's GCI condition. Finally, since the 4v4 trials were usually scheduled as the first, and sometimes the last, trials of the day, the aircrews had some expectation of the number of opposing aircraft they were to encounter before they started. As a result of all this knowledge about aircraft numbers, starting positions, and separation distances, the aircrews were able to develop specific tactics to cope with known force structures.

As we noted in the discussion on unrealistic test conditions, the aircrews could stylize their tactics because they had nothing as complicated as an overall mission to accomplish. Originally, two different test scenarios had been defined. For the neutral trials (in which neither force had the advantage of GCI information), the friendly force would be on a fighter sweep mission to clear a given area of enemy aircraft, which in turn were to intercept the aircraft that had penetrated their area, which was assumed to be near the forward edge of the battle. For the trials in which one force or the other had the GCI advantage, a fighter escort was postulated. The initial condition would allow an attacking force with GCI information to achieve as many credible surprise options as possible. The defending force, in turn, would not have GCI information but would be allowed to make use of its normal tactical procedures and equipment in order to counter surprise attacks while still performing its assigned escort mission.

The escort missions were pretested, but the aircrews ignored the escort aircraft in order to focus on keeping their losses at a minimum and on killing as many of their opposition as possible. These were the primary measures of effectiveness, and there was no measure of effectiveness for the escort scenario. Rather than defining other measures of effectiveness or establishing new test-control operating instructions for it, the JTF simply dropped the escort scenario from ACEVAL just before the AIMVAL test trials began. This change allowed the pilots to increase their success in terms of favorable exchange ratios in unlikely battle conditions. Although the JTF and the services later criticized ACEVAL for this lack of realism, it was a problem they themselves had created by not adhering to the test's design.

The pilots' perfect knowledge, their unrealistic mission, and their stylized tactics enabled them to reduce their losses. They were free to disengage to a safe area, and only one visual detection, by either force, made a trial acceptable as valid. Therefore, the number of aircraft in sufficiently close proximity to fire or the number that actually fired their weapons was often less than the number nominally involved at a trial's start. That

is, the number of participants "presented" by a force differed from the number in the force ratio and the encounter size by which the trials were analyzed and reported.

For example, in a 4v4 trial with enemy GCI advantage, the friendly force might have one flight member fire one missile and then disengage, the three other aircraft flying through the range to their "safe" area without any significant interaction. This would be, in effect, a 1v1 engagement but counted as the 4v4 trial that it was intended to be. Analysis by the F-15 aircrews and the Air Force's on-site analysis team indicated that the proportion of "mislabeled" trials was unequally distributed among the planned trial sizes and that their effect on the aggregated data was not known. ACEVAL's reports label trials by the intended numbers of aircraft, not by the numbers resulting from the aircrews' tactics. The former may have been larger than the latter, and the effect of the variance is not known.

The 4v4 trials were also characterized by less activity than the others because of perceived instrumentation problems. Although the JTF's review of instrumentation problems did not fully support this, the aircrews believed that it was more difficult for the air combat maneuvering instrumentation range to record data during trials with a lot of activity. The F-15 aircrews stated that they reduced their activity in order to increase their chances of having valid trials. Many of the complex, highly active trials--those with more aircraft, several shots, and heavy workloads for ground monitors--were declared invalid because data were lost through the aircrafts' measuring instruments, shots went unrecorded, simulations of missiles were not available although as many as eight were allowed at one time, killed aircraft could not be removed, and ground monitor communication to pilots was confused. For example, one F-15 4v4 trial took 11 attempts to complete as a valid trial. Whether or not the data collection system actually posed problems, the aircrews seriously doubted its ability to document a large-scale air-to-air encounter. They claimed to have modified their behavior and tactics in order not to overload the system, making the 4v4 trials both unrealistic and unlike the smaller trials.

Analysis problems

All the problems we have discussed for these steps of the test process make the comparison of the quantitative results of the 4v4 trials with other trials questionable. Their dissimilarity is apparent in that they were more likely than the smaller trials to be declared invalid. Approximately 30 percent of the smaller trials were invalid; 51 percent of the 4v4 trials were invalid. According to the JTF's analysis, the friendly force exchange ratios were the worst during invalid trials. Thus, it appears that the procedure of eliminating invalid trials from the data base, if it introduced any bias in the test results, did so in favor of the friendly force, especially in the trials with the largest numbers of aircraft.

Other analysis problems stem from the measures of effectiveness that were used. The JTF stressed that exchange ratios and loss rates should not be examined alone, noting that exchange ratios are an overall measure of the ability to kill but can be misleading because they do not reflect either the total number of encounters or attrition. For example, a 2v4 trial could result in the loss of two enemy and two friendly aircraft or in the loss of one enemy and one friendly aircraft: the exchange ratio (1:1) would be the same. Also, some of the big differences in aggregate exchange ratios that were reported are really the difference of only one or two kills because of the small sample sizes.

Loss rates, however, do account for attrition or, conversely, survivability, and thus they provide information not given by the exchange ratios. But loss rates have two problems that they share with exchange ratios. First, neither one of these primary measures of effectiveness was a measure of ACEVAL's overall mission success; they measured only fighter versus fighter engagements. Second, the loss rates and exchange ratios depended on probability-of-kill models simulating weapon scoring and target vulnerability. Measures of effectiveness that depend on such models are only abstract indicators of test activity. The JTF attempted to insure that they would yield valid estimates, but a number of the problems we have noted above affected the quality of the ACEVAL data so that the results from the models may be biased (app. VI, item 7).

To the JTF's credit, it analyzed data from air combat other than loss rates and exchange ratios. The proportion of "aircraft targeted" and "fired at" and targets "endgamed/intercepted" (a missile that "endgames" is successfully guided to a target, regardless of the probability of kill after intercept) were based on data generated during the test, not on the simulation models. However, since the JTF relied on loss rates and exchange ratios to express test "outcomes," these other measures were used primarily to explain them, so that a large proportion of the test activity was glossed over. For example, only about 12 to 22 percent of the attempts to simulate a missile firing resulted in a kill and thus were included in loss rates and exchange ratios (app. VI, item 8).

Reporting problems

In reporting the force ratio and encounter size findings, the JTF appropriately cautioned readers about the test's constraints but, after drawing conclusions about the causes of the observed outcomes, recommended "improvements" for the missiles and their avionics for which there were no bases in the test data. For example, the JTF attributed the finding that effectiveness decreased as the force ratio decreased to hardware problems--specifically, to limitations in the radar missile. However, since different hardware configurations had not been tested, causes rivaling the hardware problems had not been ruled out. The JTF used the percentage of friendly aircraft killed while firing the AIM-7 percent) and the percentage of friendly aircraft killed by the

target at which they were firing the AIM-7 percent) as indicators of the radar missile's long fire-to-intercept time and target-tracking limitations. Similarly, the JTF attributed the finding that effectiveness decreased as encounter size increased to "human" factors such as confusion and loss of coordination, although these were not measured directly. Thus, the JTF's report of its conclusions is largely based on inference from indirect measures, and the recommendations for improvements may be misleading.

Summary of threats to test results for the aircraft numbers objective

The problematic aspects of the JTF's assessment of the effect of encounter size and force ratio in the ACEVAL test program seriously threaten the quality of the JT&E's results. It appears clear that the problems we have discussed for the aircraft numbers objective distorted the extent to which encounter size and force ratio affected loss rates and exchange ratios. It is possible to speculate that the bias favored the friendly force overall, but it is not possible, with the available data, to be certain about the direction and magnitude of the error.

Elaboration of test objective and reported results

The objective on ground control intercept information was to determine how the presence or absence of GCI information affects loss rates and exchange ratios in air-to-air combat. Three conditions of information were tested: only the friendly force had it, only the enemy force had it, or both had it. There were no trials in which both forces were without GCI information. The effect of the three conditions was examined for seven combinations of friendly and enemy forces.

The JTF reported two major effects of having GCI information on loss rates and exchange ratios. For one, it helped the friendly forces more than the enemy. It did not significantly raise or lower friendly force losses, but it did help in killing the enemy. For the other, having GCI information decreased as an advantage against an opponent that did not have it as the number of aircraft increased. The JTF attributed this decrease in effectiveness

GROUND CONTROL INTERCEPT OBJECTIVE

JTF objective

Determine the effect of the availability of ground control intercept information.

JTF conclusions

"GCI's [ground control intercept's] primary effect was on [enemy] loss rates; when [friendly] had GCI, [enemy's] loss rate was significantly higher; there was no significant effect on [friendly] loss rates with or without GCI. . . . In the 4v4 trials, the GCI effects on [enemy] or [friendly] loss rates were negligible."

Figure 46
The Effect of the Advantage of Having
Ground Control Intercept Information

to the increase in the complexity of the communications that are required in the larger trials. (Figure 46 shows the effect on the F-14 and the F-15 of having ground control intercept information, by exchange and force ratios; see II.B.13, p. VI-22).

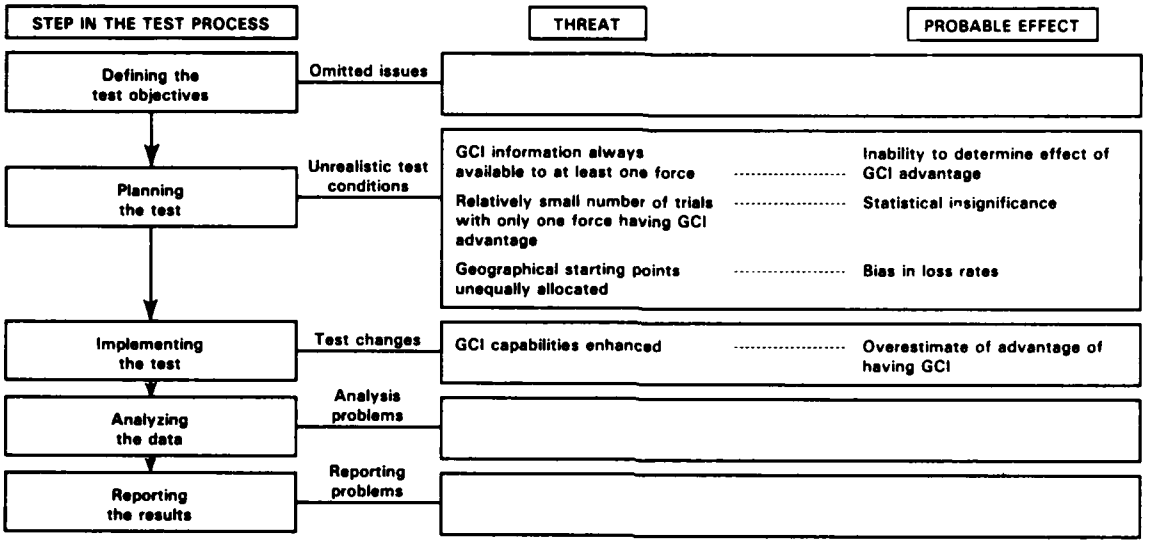
Threats to test quality

The threats to the quality of the results on ground control intercept information and their possible effects are summarized in figure 47. There were threats at the steps of planning the test and implementing it.

Unrealistic test conditions

ACEVAL's feasibility study stated that fully testing encounters within visual range would require not only the three conditions in which at least one force had GCI information but also the condition in which neither force had it. However, the final test matrix (figure 41) did not include any trials in which neither force had GCI information, and no explanation was given for dropping this condition. The JTF test plan called for a limited number of trials with jammed communications if time and resources were available at the end of the test, but these trials were not flown. No baseline data are available for making comparisons between what happens when at least one side has GCI information and what happens when no side has it--a condition realistically to be expected during combat.

Figure 47
Threats to Test Quality: The Ground Control Intercept Objective



There were two other problems with respect to realism. Each force was without GCI in only 26.7 percent of the trials. As a statistical sample, this is too small, since it diminished the opportunity to measure what can realistically happen, especially in 4v4 trials, where, as the JTF had noted, larger numbers of aircraft make for more complex communications problems. Further, the geographical starting points on the testing range were not equally allocated among the three conditions--there were 18 possible points for the neutral trials (those in which neither side had the advantage) but only 10 for the trials in which one side had GCI information but not the other. Since starts in which friendly and enemy forces were in close proximity to each other were used only in trials in which one side had the advantage, and since the number of these trials was relatively small, the aircrews were able, unlike in real combat, to determine very quickly which situations were disadvantageous or threatening and to disengage to a safe area. The reported loss rates may be underestimates.

Test changes

A number of situations involving GCI capabilities that had originally been planned for were not tested in ACEVAL, so that the effectiveness of the GCI controllers was much greater in the test than can be expected in normal training or combat. First, there were never more than four aircraft on a side, never more than a total of eight, on the range. Actual combat might involve 20 aircraft or more. Second, the minimum altitude was relatively high; had it been lower, at feet, the aircrews could have flown underneath the level at which they would need GCI

coverage. The predominant use of GCI information meant that what happens when aircrews fly low enough not to need it was not tested. Third, the GCI controllers knew the signatures of both the friendly and the enemy aircraft, whereas in combat what is more likely to be available is merely radar reflection or paint, making it more difficult to establish the identities of friends and foes with certainty. Fourth, the debriefings that made use of displays from the air combat maneuvering instrumentation range provided exceptional learning opportunities. Fifth, the jamming of communication that can be expected to hinder GCI capabilities during combat was not included in the ACEVAL trials.

Not allowing for all this reduced the uncertainty about the intentions and possible movements of the opposition. This relieved the GCI controllers from having to perform under the difficulties that would be posed by having more limited information in combat.

Summary of threats to test results for the ground control intercept objective

The efficiency of the GCI controllers during ACEVAL was tested under highly favorable conditions. Although some of the environmental features of the test could not have been changed, no attempt was made to restrict the information that was available to the controllers, or that they made available to the aircrews, in a way that would be more representative of combat. The effect of the exaggerated availability of information, from the situation of the controllers and from the situation of having no trials in which neither force had GCI information, cannot be determined. However, given the results of the 4v4 trials shown in 46--those that came closest to portraying the complexity of air combat realistically--the JTF's conclusion that "normal" GCI does not increase the effectiveness of friendly forces is not warranted.

Elaboration of test objective and reported results

The primary purpose of the objective on aircraft type was to provide an indication of whether the effects on aircraft numbers

AIRCRAFT TYPE OBJECTIVE

JTF objective

"For each initial condition [start point and ground control intercept status], determine whether the outcome varies with the type of aircraft (F-14/15)."

JTF conclusions

"There were significant differences between the F-14 and F-15 loss rates overall; however, only in the lower force mixes were there substantial F-14/15 differences in exchange ratio. These were caused by the limitations of the F-14 fire control system and the hotter F-14 [infrared] signature at idle thrust caused by the TF-30 engine Mach lever interface."

Figure 48
The Relation Between Force Ratio and Exchange Ratio for Two Aircraft Types (Normalized)

of the different GCI information conditions are the same for different types of friendly force aircraft. The JTF reported that the F-14 and F-15 were able to cause approximately the same number of enemy losses but that the F-15 exchange ratio was higher since the F-15 aircrews suffered approximately
than the F-14 aircrews. As shown in figure 48,
exchange ratios for the F-15 occurred in trials in which the friendly force outnumbered the enemy force and in the smaller trials in which the two sides were even. (See II.B.15, p. II-16.)

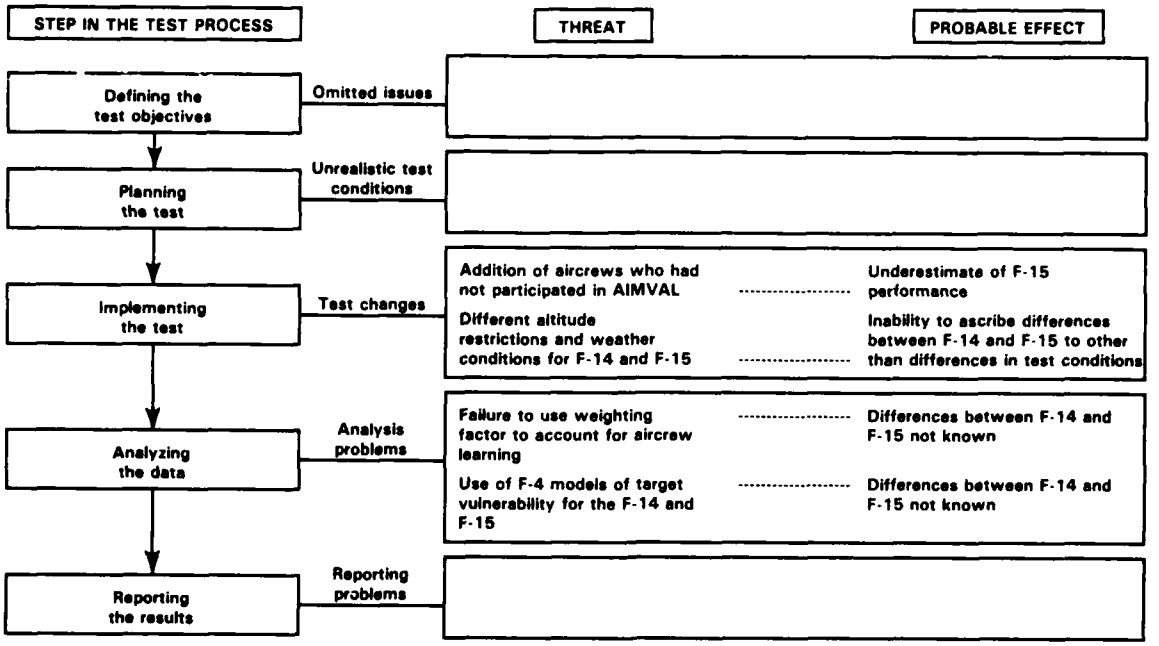
Threats to test quality

Threats to the quality of the ACEVAL results for the objective on aircraft type are summarized in figure 49 (on the next page). There were threats at the steps of implementing the test and analyzing the data.

Test changes

The ACEVAL test design established a comparison base appropriately by setting up equal numbers of F-14 and F-15 trials to be flown under identical conditions of force ratio, encounter size, and GCI advantage. The Eagle Eye II optical aid for the F-15 gave it a visual-aid advantage similar to that provided by the television sighting unit to the F-14. However, it is inappropriate to make unequivocal comparisons between the two types of aircraft or

Figure 49
Threats to Test Quality: The Aircraft Type Objective



to combine the two data bases, because they were biased by the learning of the aircrews and by an inconsistent application of test-control operating instructions for weather and altitude.

In order to expedite the Air Force test trials, F-15 and F-5E pilots (who had not participated in AIMVAL) were trained from July 11 through August 1 and integrated with their forces. The new F-5E aircrews showed no measurable difference from their counterparts who had had AIMVAL experience, but the test data do reveal learning that is attributable to the two new F-15 pilots (app. VI, item 9). When the JTF removed the new F-15 aircrew trials from the analysis, the success rates of the aircrews who had had experience in AIMVAL were the same throughout ACEVAL; the success rate for the new aircrews was one third lower for the first half of ACEVAL but no different thereafter. However, the JTF aggregated all the F-15 trials without using any weighting factors to account for these effects. As a result, the data for the F-15 include effects from learning while the data for the F-14 and F-5E do not.

Overall, ACEVAL was flown in good weather, but the interpretation of "minimum" weather was left to the aircrews. The F-15 pilots interpreted it more stringently than the others and flew in weather that was better than that the F-14 aircrews flew in. The enemy aircrews and the Navy's on-site analysis team noted that the

1v1 and 1v2 trials were hampered because engagements occurred below the overcast, whereas the 4v4 trials were favored by being above the overcast (app. VI, item 10). Engagements below an overcast made visual detections and shot opportunities easier, especially for the enemy pilots, given that the F-14 was much easier to detect when it was flying against a cloud background. Engagements above an overcast gave the friendly forces an advantage because the was more effective at the greater range and in the absence of ground clutter. The was not as effective, because the cloud background made it more difficult to separate the infrared signature of a target from the thermal clutter of the background. As a result, the F-14 trials might have had different outcomes (better or worse) in different weather, but the testing did not control for this, and the effect of the pilots' interpretations is not known.

Another difference in the test conditions for the two types of aircraft was that the F-15 was permitted to fly at lower altitudes. A minimum altitude had been set for both aircraft at feet above mean sea level, an altitude that accommodated the abilities of the air combat maneuvering instrumentation system, but instrumentation problems were more pronounced for the F-14 at this altitude because of its low-altitude tactics, so the level for the F-14 was raised to feet. It remained at feet for the F-15. Without a control for the difference, the data cannot be compared.

Analysis problems

The simulation models for estimating aircraft losses used the vulnerability of the F-4 as a target instead of the vulnerability of the F-14 and F-15. The JTF provided no analysis or discussion of the differences, so that the ACEVAL results are difficult to interpret. Differences between the F-14 and the F-15 may, in fact, be the result of improper modeling rather than operational performance.

Summary of threats to test results for the aircraft type objective

Data for appropriately comparing the single-seat and two-seat aircraft were not collected. The JTF did not control for differences in aircrew learning, weather, or altitude. The JTF compared and then combined the F-14 and F-15 data, attributing observed differences between the two aircraft to differences in avionics that might be more accurately attributable to how the F-14 and F-15 pilots interpreted and followed the test-control operating instructions. In addition, the differences that were found may actually have been more or less, by some measure that cannot be known, because the models that were used for determining the probabilities of kill were based not on the F-14 and F-15 but on the F-4.

COMBAT ELEMENTS OBJECTIVE

JTF objective

Determine the effects of the primary control variables of encounter size, force ratio, and ground control intercept conditions on tactics, hardware, aircrews, and "key" activities such as visual and radar detection, visual identification, firing, kills, and losses. In addition, identify variables that affect encounter outcome and determine how the effect of each is related to encounter size.

JTF conclusions

On encounter size, "The trends apparent in ACEVAL were that advantages important to [friendly] in a 1v1 fight, such as superior weapons systems, GCI and attainment of key activities prior to the opposing side, decreased as the numbers increased for a fixed force ratio. . . . As the numbers increased, the human element of confusion and reduced coordination seemed to be increased."

On force ratio, "when [friendly] was facing superior numbers, there were always free [enemy] fighters; however, [enemy] did not have the sole advantage in the visual arena as he was also within [friendly] [infrared] or radar lethal envelope. The net effect was that [friendly] was unable to consistently reduce the odds to parity or better, prior to arriving within [enemy's] visual range."

On ground control intercept conditions, "The primary utility of GCI for [friendly] was for initially locating the [enemy], particularly in altitude, (in the smaller force sized engagements) and in reattack; without GCI there were very few reattacks. . . . [I]n the 4v4 situation, close GCI control began to contribute to the confusion problem."

On other variables, "The primary control variables of force size, ratio and GCI accounted for only 10-20% of the variation between individual trials. . . . There were many other factors that influenced the losses in individual engagements. However, number and GCI were the primary factors in aircrew implementation of decisions regarding tactics, coordination, firings and disengagements."

Elaboration of test objective and summary of threats to test results for the combat elements objective

The JTF examined how the independent or primary control variables (figure 42) affected various combat elements such as tactics, hardware, aircrews, and key activities of the battle (such as detecting a target first or firing a weapon first). The data that the JTF collected included aircraft position, velocity, and acceleration; weapon targeting, interceptions, and kills; and reports from the pilots on their tactics.

The JTF found that increasing the number of aircraft decreased the utility of certain advantages that the friendly forces had in 1v1 fights. These advantages were their superior radar and avionics and their ability to detect and identify targets earlier than the enemy forces could detect or identify them. In addition, the JTF reported that increasing the number of aircraft in a trial confused the aircrews and reduced their coordination.

As for adverse force ratios, the JTF reported that the friendly forces were always faced with some enemy aircraft that

were free to maneuver. This reduced the ability of the friendly forces to kill enemy aircraft and survive the task.

When the friendly forces had ground control intercept information, the JTF found, their aircrews were able to use this information to their advantage for initially locating the enemy. More importantly, this information enabled friendly aircrews to attack more than once before disengaging from the trial or before it ended. However, close GCI control confused the aircrews in the 4v4 situations. While the JTF reported that variables other than the independent ones accounted for 80 to 90 percent of the variation between individual engagements, it also noted that the number of aircraft in an encounter and the presence of ground control information were the primary influences on the aircrews' decisions during the test.

The quality of the test results on the combat elements objective is predicated on the quality of the data for addressing the aircraft number (in terms of force ratio and encounter size), the aircraft type, and the ground control intercept objectives. In other words, how well the independent variables were tested affected the ability to determine how these variables affected other elements of combat. As we have demonstrated in this chapter, various problematic aspects of assessing encounter size and force ratio threatened the quality of the test results and probably biased the data in favor of the friendly forces. In addition, the efficiency of the GCI controllers was tested under highly favorable conditions. Consequently, the JTF's conclusions about how the independent variables affected other aspects of combat are influenced by the potential biases of the test conditions that we have discussed in relation to the other objectives of the test.

It should be noted, however, that despite the quantitative limitations of the test data, the JTF did report, in a separate volume of its final report, discussions between the members of both enemy and friendly forces on their observations about ACEVAL and the tactics they used in the test (see our appendix II.B.12). This allows a qualitative assessment of how the independent variables affected the aircrews and the trial outcomes.

SUMMARY OF QUALITY

ACEVAL was an ambitious undertaking and the first major program to use flight tests with several aircraft simultaneously on the recently acquired air combat maneuvering instrumentation range, designed to inform pilots in flight about the progress of combat and to record the data necessary for a thorough analysis of its results. Despite these auspicious features, controllable aspects of ACEVAL were not controlled, creating the problems of quality that we have described in this chapter.

While we have focused on aspects of the test program that would have improved the test's quality had they been accomplished

Figure 50

Summary of the Quality of ACEVAL Results

The test set out to

> determine how the outcome of air combat depends on the number of aircraft engaged on each side under various conditions.

A detailed test planning process took place that

> identified the issues of air-to-air combat and the requisite test parameters and resources and that specified the instrumentation needed to provide simulated combat information to pilots in flight and to record test data.

Various events and decisions led to

> relatively few 4v4 test trials, more favorable test ground control intercept information than in combat, test equipment and instrumentation favorable to the friendly forces, and differences in the implementation of the F-14 and F-15 trials.

Consequently, it is appropriate to state about the test's results that

> --the results for the encounters with larger numbers of aircraft on either side may not be comparable to the results for the 1v1 encounters;
> --the data on the contribution to combat effectiveness of information on aircraft position and direction relative to identified targets may be overstated;
> --the conclusions on the effectiveness of the Navy's F-14 and the Air Force's F-15 may be overestimated;
> --the perceived differences between the F-14 and F-15 trials may not be attributable to differences in the aircraft and their equipment.

differently, we have touched in passing on a number of other constraints on the quality of the test's results. These included the unique features of the test range and the training, the limitations of the F-14's long range missile, and the absence of ground-to-air countermeasures threats. In figure 50, we summarize the effect of the threats to the quality of the test's main objectives.

THE USEFULNESS OF THE TEST RESULTS

The JTF's intended use

If test results are to be useful, they must be relevant, sufficiently high in quality, well presented, and timely. Regarding relevance and quality, OSD had expected ACEVAL to provide empirical data for extrapolating from one-on-one to multiple air combat encounters in order to fill a gap in what is known about how aircraft numbers affect air combat. Although its scope was limited, ACEVAL could have yielded important information about encounters within visual range and established baseline data for understanding the results of tests designed to systematically vary the avionics and the weapons. However, it is doubtful that ACEVAL provided valid answers to the general questions that were posed when the test was nominated or the specific questions that were posed when the scope of the test was more narrowly defined. The test's many technical inadequacies lead us to question the utility

of the ACEVAL data for developing either generic models for analyzing air combat with large numbers of aircraft or specific models for simulating the test's outcomes with other weapon systems. Therefore, we believe that the requestor's needs for empirical data for analytical purposes were not fulfilled.

As for the presentation of ACEVAL's results, the JTF provided appropriate caveats in its reports about the limitations of the test program, cautioning readers that

> "Descriptive, derived and postulated ACEVAL results should be interpreted with caution. Although many data quality controls were used, the nature of the test, data collection procedures and the trial validation process introduced various uncertainties into the ACEVAL statistics. Standard analysis methods wre used to accommodate these uncertainties where possible; however, the derived results were still affected, and the danger of misinterpretation is high." (II.B.13, pp. VI-2 to VI-3)

The JTF's presentation of the results went so far as to include the perceptions of the aircrews and a summary of the raw trial data, as well as the JTF's empirical and subjective findings on each test objective. It was fairly balanced in detail and emphasis. The notable exceptions were failing to put in the main report the details of the advantage that the friendly forces had beyond visual range--these were tucked away in an appendix--and making a statement of recommendations for improving missiles and the avionics when these had not been tested. Four of the five volumes of the final report were classified secret, limiting their distribution. The only unclassified material was a discussion of the management lessons that had been learned. We are not aware of any unclassified summaries of the report for general distribution.

Although the time proposed for the test slipped from 1974-76 to 1974-79, "timeliness" was not affected, since no critical time constraints had been imposed. In fact, AIMVAL's coming first and using many of ACEVAL's basic design elements gave the services and the Congress acquisition-related information sooner than had originally been planned.

Other uses

ACEVAI was proposed as the initial test of a series of operational multiple air combat tests. Its scope was purposely limited, in the expectation that other tests using ACEVAL's procedures would later address a variety of specific and practical issues. However, AIMVAL is the only other test like ACEVAL that has been conducted. The AIMVAL and ACEVAL data being the only operational test data available on multiple aircraft combat, they have been and continue to be used extensively, despite their limitations. Instances of other, largely inappropriate, uses of the test results are described in appendix VI, item 11.

The references that have been made the most often to the AIMVAL and ACEVAL results appear in the context of the debate on DOD's acquiring better versus more weapons. The test data have been cited to support opposing viewpoints. OSD and the Air Force have used the data to develop analytic models on larger numbers and different weapons. These uses may be appropriate for deciding what hypotheses to use for future testing. They are not useful for providing information for any other kind of decision, because the ACEVAL data reflect merely the idiosyncrasies of this test's design and implementation.

ACEVAL's results have been more usefully applied by the Air Force and the Navy for gaining insight into tactics and training. Further, the Air Force put ACEVAL's lessons to use when designing the "manned" simulation phase of the operational utility evaluation of AMRAAM, a Navy-Air Force advanced medium-range air-to-air missile.

Insuring that knowledge finds its way into the preparation of future testing, in the hope of avoiding the pitfalls of earlier efforts, is an excellent way to increase the usefulness of JT&E. However, beyond its utility for AMRAAM, what was learned about testing in ACEVAL has not been put to use in other operational tests, because none have been conducted. A current plan for conducting another air-to-air combat JT&E would substitute simulation altogether for real flight with live pilots.

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

The need for joint military testing and evaluation was acknowledged in 1970, when the Blue Ribbon Defense Panel recommended that a Defense testing agency be created with the responsibility of conducting the overview of all Defense testing and evaluation and of conducting tests and evaluations that span the armed services. DOD did not create a testing agency but did acknowledge the need for joint tests and evaluations: in 1971, the task of insuring productive JT&E's was assigned to an existing DOD office.

The three JT&E's we examined for this report were all complicated operational tests on a wide range of military issues conducted between 1977 and 1980. The IIR Maverick JT&E addressed many aspects of employing the IIR Maverick missile system in two battle scenarios and was intended to involve the Air Force, the Army, and the Navy. The TASVAL JT&E addressed close air support in a joint environment with more than 100 instrumented players from the Army and Air Force. The ACEVAL JT&E addressed important issues regarding the structure of forces in air-to-air combat, using Air Force and Navy fighter aircraft to simulate "friendly" and "enemy" forces. DOD has, indeed, developed the means of conducting JT&E, and the JT&E's it has conducted have not been simple.

Given the many complex issues, the quality of each JT&E depends on how well the various steps of the test process are performed. When issues are more clear-cut, it is easier to formulate a test's objectives and, consequently, to design the test. In other words, the test process is cumulative. This means that DOD's JT&E program necessarily has various limitations. While acknowledging the noteworthy attempts that DOD has made to conduct productive JT&E's, we believe that improvements are needed to eliminate or at least minimize the effect of various events and decisions, made during the test process, that seriously threaten the quality of JT&E's results. Accordingly, we have organized our remarks in this chapter around, first, the questions that we raised in the main body of our report about the way DOD's joint tests and evaluations are conducted, from their initiation to the use of their results, and, second, the recommendations that we believe follow from our findings about the test-and-evaluation process.

CONCLUSIONS

How independent is the DOD organization that is responsible for conducting JT&E from other DOD organizations that have vested interests in JT&E's results? A recent statute provides that a civilian Director of Operational Test and Evaluation in DOD report directly to the Secretary of Defense as OSD's principal advisor on operational test-and-evaluation matters. This director is to be DOD's senior operational test-and-evaluation official. Some

of the functions of this office will be similar to those of the DDT&E, but what the relationship between the two offices will be is not yet clear. If the joint test-and-evaluation program is thus removed from the DDT&E's responsibility, it will become organizationally independent of the Under Secretary for Defense Research and Engineering, where it is now responsible to OSD's weapons developer organization.

In the past, tests were managed, carried out, and even partially funded by the separate services. The JT&E program greatly relied on them for test sites, instrumentation, and equipment; for managers and other personnel to carry out the tests; and for operations and maintenance funds to pay for testing activities. Consequently, the JT&E program has depended to a large degree on cooperation. It is not yet clear whether creating a new office for the direction of operational testing and evaluation will make the JT&E program independent of others for its resources and capacities.

Who requests joint tests and evaluations and why? The purpose of JT&E is to ask questions about the ability of developmental and deployed weapon systems to perform their intended missions when two or more services are engaged jointly in combat. We found, however, that the most frequent requestors of the 13 JT&E's that had been completed at the time of our review were not those groups with the greatest responsibilities for joint military planning and performance. We found that only 3 tests actually involved the services in joint operations, although at least two services participated in all 13.

The relative absence of the JCS as a requestor of tests is conspicuous. According to DOD Directive 5000.3, the JCS is DOD's main proponent for joint procedures and the interoperability of deployed forces, and it should have questions whose answers come from JT&E. While the JCS is potentially the biggest user of JT&E, it has not accepted the JT&E program as a way of examining the structure and combination of military forces. We found that only two of the JT&E's conducted between 1972 and 1983 were requested by the JCS. The JCS has stated its belief that field exercises are more valuable and yield more timely information than the quantitative data that come from JT&E. The DDT&E began using new procedures in 1981 for greater and more systematic participation by the JCS, especially in the nomination and selection of JT&E's. We observed that the JCS does participate in this more formal process but could not determine to what extent.

Although most of the 13 completed tests had several objectives, each of them focused predominantly on a single goal. Three were conducted primarily to provide data for weapon-system acquisition decisions, 4 sought to establish whether the hardware or system design requirements or the operational capabilities of deployed or developmental systems could be met, 4 evaluated techniques for improving testing methodology, and 2 tried to determine the utility of procedural or technical concepts for existing

or developmental weapon systems. We found no evidence of an overall agenda or a strategic plan for insuring that major issues of joint importance would be addressed by the JT&E program. The DDT&E's 1981 procedures are making the nomination and selection of joint tests more systematic, but the nominations are still ad hoc.

Are JT&E problems defined to include critical operational issues? The IIR Maverick, TASVAL, and ACEVAL tests were all designed to address critical operational issues, but many important issues were not included or even acknowledged in them. Clearly, all issues cannot be included. If they were, the tests would become too complex. Our case studies show, however, that the missing components were in some instances so integral to the overall question being addressed that their absence seriously damaged the usefulness of the test results. Furthermore, some of the missing components could have been included by incorporating results from previous testing--that is, a cumulative or building-block approach could have been used but was not.

We also found that the focus in the three tests we examined was more on the hardware aspects of operation than on human abilities and performance. As a consequence, the tests tried to demonstrate how well a weapon could meet its technological specifications but gave little attention to how well the operator of that weapon could conduct a mission with the available supporting systems. Furthermore, when a test's objectives were stated in terms of "kills" and "losses," the test results constituted a prediction of performance that was based on computer models rather than "combat" events with live participants.

The omission of critical operational issues is well exemplified by the IIR Maverick test. According to Air Force doctrine, the ability of pilots to distinguish enemy ground forces from friendly ground forces is a critical operational factor in close air support missions, but it was not included, nor discussed as an assumption, in the IIR Maverick test, even though the test's objective was to examine the pilots' ability to provide close air support. Similarly, a critical operational factor in air-to-air combat is the mission objective, but no specific mission objectives were defined for the ACEVAL test: the fighter pilots needed only to survive, without having to consider how to accomplish any specific mission. ACEVAL's lack of a specific mission objective was, however, discussed explicitly in the test's report.

JT&E's lack of continuity and cumulativeness was most evident in TASVAL. For example, an earlier JT&E, Close Air Support Command and Control, had demonstrated that the shortest average times for aircraft arriving at the forward edge of the battle area in response to call for close air support from ground forces may be minutes after leaving the base for rotary aircraft but minutes for fixed-wing aircraft, but the TASVAL close air support scenario did not simulate this difference. Instead, the JTF assumed unrealistically that they would arrive simultaneously at the forward edge of the battle. Moreover, the possibility

that the friendly and enemy ground forces would be mixed, and that the pilots would have to differentiate between them, was reduced or eliminated because the aircraft arrived at the battle shortly after ground activity had begun. The results of another operational test, the Joint Attack Weapons Systems Tactics Development and Evaluation, could have been but were not used in defining TASVAL's objectives regarding the joint air attack team's tactics for close air support.

In the IIR Maverick JT&E, we found no evidence that results from earlier electro-optical Maverick operational testing were considered in the definition of the issues that the IIR Maverick test was to address or in the JTF's discussion of its results. We did find, however, that the test conclusions state that the IIR Maverick missile has better capability than the electro-optical Maverick missile.

We found that JT&E emphasizes the machine rather than the human aspect of weapon systems. For example, in the IIR Maverick JT&E, factors such as a pilot's ability to use the missile to acquire and lock on to a target were examined more closely than the pilot's ability to find the target area while flying at low altitude and in poor weather. In the TASVAL JT&E, factors such as the ability of the joint air attack team to kill the enemy were examined more closely than the ability of the members of the joint air attack team to coordinate their attack efforts.

Finally, stating a test's objectives in terms of an assessment of "kills" and "losses" makes its results dependent on models and on a weapon system's technical performance. Tests cannot fully assess the killing and loss that happen in combat. Instead, data must be manipulated by using computer models that simulate the conditions under which artillery should be successful in warfare. The attrition and effectiveness objectives in TASVAL were stated in a way that required computer simulations of how well the attack aircraft could survive and how effective the missile could be in killing enemy ground forces. Similarly, the major objectives about the effects of aircraft numbers for ACEVAL and the survivability objective for IIR Maverick necessitated the application of models.

All this stands in contrast to the objectives of the IIR Maverick test that focused on the ability of the pilots to find the target area and to those of ACEVAL that focused on a pilot's making the first identification of the enemy or firing the first shot. Making test results dependent on models can make them misleading, because the test data that are used as inputs do not always meet a model's assumptions and because models themselves are not always valid indicators of system performance. Test results that derive from the activity of the participants give better estimates of combat performance.

Do the design and implementation of joint tests generate reliable and valid data about the operation of weapons systems,

their limitations, and the concepts of their employment? To say that test results are reliable is to imply that repeated testing under the same planned conditions would yield roughly the same results. Reliability implies consistency, not that results are correct.

The test designs we examined generally spelled out what was to be done in considerable detail. The number of trials for selected combinations of independent variables (both controllable and uncontrollable) were specified. Mission scenarios and test procedures were generally set forth with particularity about how trials were to be executed. Planned measurement procedures were necessarily sophisticated, given the complexity of what was being measured. Consistency in recording the data was controlled for by procedures that deleted data sets from whole trials or parts of trials when implementation problems affected them. Test designs were weak regarding reliability when they called for too few trials for addressing critical objectives.

The JT&E's we reviewed do not allow us to make a general statement about whether the designs were followed by the kind of implementation that is necessary to produce reliable results. Uniformity in the execution of missions from trial to trial, a minimum of irrelevant variation in the establishment of environmental conditions, and few instrument errors in measurement all help insure reliability. We found that such implementation factors were conducive to the production of reliable data in some tests but not in others. In TASVAL, for example, the data collection was so burdened with problems that more than 50 percent of the data were missing for some aspects of the test. However, details about these problems in the test's implementation were not always available, and we can make no judgments about their overall effect.

The extent to which JT&E results are valid is a different question. To say that test results are valid is to imply that they measure what they purport to measure. A valid result in JT&E is one that accurately predicts the ability of the armed forces to perform in combat. From our review of three JT&E's, we believe that the validity of many of their results is seriously doubtful.

This belief is based on three main considerations. Unrealistic test conditions were laid out in most of the test designs, and realism broke down further during test implementation. Competing explanations for the tests' results were not eliminated or explained. Some of the steps of the data analysis were questionable. The consequence of these threats to test quality is that many of the test results overestimate the combat capability that was being tested. That is, the tests did not measure what they purported to measure, and predictions based on them may overestimate the likelihood of success in combat.

Unrealistic test conditions were common among the three JT&E designs. Examples of what could have been designed with more

realism include the use of a single, small target area in the IIR Maverick test, the absence of smoke and fire and dust on the TASVAL battlefield, and the availability of a safe disengagement area for the aircrews in ACEVAL. Lack of realism during test implementation may have been difficult to control. For example, in the Maverick scenarios the points at which aircraft were to ascend before beginning an attack were not varied as much as had been planned, with the consequence that pilots were able to gain a familiarity with the battlefield terrain that they would not have in warfare. Similarly, in ACEVAL, the aircrews were able to acquire an improbable knowledge of opposing force numbers, locations, and intentions before "combat" began.

In some instances, competing explanations for a test's results were not examined. In TASVAL, for example, many results were explained as stemming from variations in the composition of the strike forces or the terrain, but other variations, as in the time of day when trials were conducted, might also have explained these results.

As for the analysis of the test data, we were often not able to tell the direction in which the results are biased, although it was clear that the results are probably biased one way or the other. One example of the difficulty is the use of questionable procedures in the analyses for adjusting for missing data. Another is the use of questionable assumptions in the computer simulation models that were used to produce test results from data collected during testing.

Overall, we believe that the test results from the three JT&E's we examined are possibly reliable but often doubtfully valid. The implication is that potential operational problems may not have been revealed by the tests. The operational problems that did appear should be taken seriously, however, because we cannot tell whether they came from random error (in which case the results may be reliable) and because we judge the bias in the results as likely to lead to underestimates of the seriousness of the problems.

Do the joint test-and-evaluation results that are reported accurately reflect the data that are collected? We found that, in a number of instances, the results presented in the JT&E were not an accurate reflection of the data that were collected. In some reports, appropriate qualifications of the results were lacking.

For example, important details of how well the IIR Maverick system performed in poor weather, under conditions of low thermal contrast, and under threats from enemy air defenses were minimized by general statements about overall "success." Data that were available and indicated a lack of success in the test under the more demanding conditions were omitted from the final report. The presentation of ACEVAL's results was fairly balanced in terms of detail and emphasis, with the notable exception that details of

the important beyond-visual-range advantage that friendly forces had over the enemy were not presented in the body of the report but were tucked away in an appendix. Stressing the positive in overall summaries of results and presenting negative results only in appendixes, if at all, makes some JT&E reports misleading.

As for the appropriate qualification of results, the ACEVAL report did provide caveats to indicate the test constraints that had affected the test results, and the IIR Maverick and TASVAL reports discussed some of the constraints but did so in a way that was not integral to their presentation of results. For example, the countermeasures results that were reported for the IIR Maverick were not appropriately linked with statements about the problems in implementing the countermeasures trials, which leads the reader of the IIR Maverick report to believe that countermeasures had no effect on the operation of the missile system while the fact is that they were not adequately tested.

Do the conclusions and recommendations that are reported accurately reflect the test-and-evaluation results? We found that the conclusions and recommendations of the joint test forces were not always supported by the test's results. For example, the conclusions about TASVAL were presented in terms of the differences between the two valleys of the test site even though the test was not designed to examine their effect. In other words, the time of day, the amount of dust in the air, and the type of defense the enemy used depending on terrain were not controlled for, so that observed differences in operational effectiveness could, in fact, be attributable to some unknown degree to a difference in the valleys. The conclusions in the TASVAL report that attribute differences in results to valleys cannot be supported because the valleys were not controlled for independently as a variable.

Similarly, the IIR Maverick report concluded that the pilots detected targets easily, that their ability to acquire valid targets improved with practice, and that the Pave Penny cueing aid was valuable to their success, but these conclusions were not supported with test data. The JTF also concluded that the ranges at which the pilots employed the IIR Maverick were "acceptable" without providing any evaluative criteria. When we compared these ranges with those that had been presented as standards in the systems acquisition report, we found that the ranges that were used in the test were not acceptable.

We also found that the JTF's recommendations were sometimes unsupported by and sometimes contradicted the JT&E results. For example, the IIR Maverick JTF reported that pilot workload was not a problem but went on to recommend four ways to reduce pilot workload. None of the ways had been examined in the test. The ACEVAL JTF made recommendations for missile and avionic improvements without having tested the alternatives.

Do the reports of the results address the concerns of the people who requested the JT&E's? For test results to be useful,

they must be relevant to the requestor's need for information. We found that the IIR Maverick report was the most relevant and that the TASVAL report was the least relevant of the three. Our judgment is based on our consideration of the omission of critical issues, the absence of a basis for comparing test results, and the lack of test conditions representative of projected combat environments.

Omitting critical issues lessened the usefulness of some of the test results. Tests that omit components that are integral to the overall question being addressed yield results that are not relevant for making sound decisions. Thi happened with the identification of friend and foe in the close air support scenarios in the IIR Maverick JT&E and with the definition of a mission objective in the ACEVAL JT&E.

Omitting essential comparison groups or baseline data lessened them further in their usefulness for deciding between alternative strategies, systems, and tactics. For example, TASVAL was to test the effectiveness of close air support at the forward edge of the battle area, but this was not possible because no baseline data were collected for a comparison with the effectiveness of ground forces without any air support. In addition, both offensive and defensive ground scenarios were planned, but only the defensive scenario was used during test trials, further limiting the data for comparative purposes. This meant that the Secretary of Defense who had requested the test did not get answers to his questions about the structure and combination of forces.

Similarly, in the IIR Maverick JT&E, the objective of comparing the workload of single-seat and two-seat aircraft was not achieved as planned since only single-seat aircraft were used. Furthermore, the JTF erroneously concluded that there "was no problem" with using the single-seat aircraft for the IIR Maverick missile, despite the lack of comparison-based affirmative evidence and the fact of some evidence to the contrary.

The ACEVAL JT&E attempted to determine how encounters between aircraft differ when the numbers of aircraft, the availability of ground control intercept information, and the type of aircraft are varied. Unfortunately, these comparisons were limited because too few of the trials with larger numbers of aircraft were included in the final test design matrix, no trials were run in which both friendly and enemy forces were without ground control intercept information, and the test conditions under which the two types of aircraft were flown were not the same.

Finally, making the tests dependent on their particular sites and instrumentation diminished their usefulness in representing projected combat. Given the limitations of currently instrumented test sites, it is not surprising that the JTF's must settle for what is available rather than for what would be best. With TASVAL, however, the test site was chosen because of concerns about the airspace and the range measurement system rather than

about how well the test results could be generalized from the test site to the central European environment being simulated. That the hot, dry climate and the desert and mountain terrain of Ft. Hunter Liggett, California, could not simulate the climate and terrain of central Europe was well known, yet the purpose of determining "aircraft losses and target kills . . . in a simulated heavily defended central European environment" was never changed, even though other aspects of the test's objectives were modified to make them more attainable given other constraints.

Furthermore, in TASVAL, the realism of a test scenario with smoke and fire was discarded, although it had been performed previously in the IIR Maverick JT&E, because of the problems it might have created for gathering data about attacks on targets by means of lasers. The JTF did not examine alternatives to the lasers that would have permitted more realistic battlefield conditions and, thereby, would have reduced the bias that favored the electro-optical Maverick and the TOW missiles that were used in TASVAL. In addition, the success of the lasers as an instrumentation device proved questionable, so that the sacrifices that were made to obtain laser-pairings on the battlefield may have been in vain. This, however, could not easily have been foreseen.

The IIR Maverick test was not as dependent as TASVAL on elaborate instrumentation requirements, but it was constrained by having only one small test area. This made the evaluation of target area acquisition questionable, because the pilots knew where that area was. Nevertheless, IIR Maverick JT&E results were the most relevant. The requestor had stated specific questions about the missile system's operational effectiveness, and the JTF attempted to address all but one. The exception was the single-seat versus two-seat aircraft problem in employing the missile. The omission meant that the relative workloads for the aircrews could not be determined. Most of the rest of the requestor's concerns were addressed in the JT&E's report, although some were presented inaccurately.

The TASVAL results were the least relevant to the needs of the JT&E's requestor. The test could not answer the questions that had been asked about combat operations in central Europe, because the test was conducted in the dissimilar climate and terrain of California on a "battlefield" devoid of smoke and fire. The test could not address the issues of weapon systems acquisition that had been raised, because the test ignored the alternatives. The factors of the structure and the combination of forces that had been questioned were not examined in the test.

Among the three, ACEVAL came closest to being relevant but this was because its purpose had been stated in only a general way by the requestor. Since looking at all aspects of air combat between many aircraft was judged infeasible, the scope of the test was narrowed to air combat within visual range, for which little operational data were then available, in the anticipation that subsequent and similar operational tests would address the other

issues. However, while the test was planned and implemented, comparative design features were dropped and unique conditions were permitted, making the results specific to that test and reducing their relevance for general analytical purposes. Therefore, while ACEVAL produced operational air combat data that had not been previously available, it did not produce the type of information that can be used in computer models to estimate what happens in a variety of air combat situations.

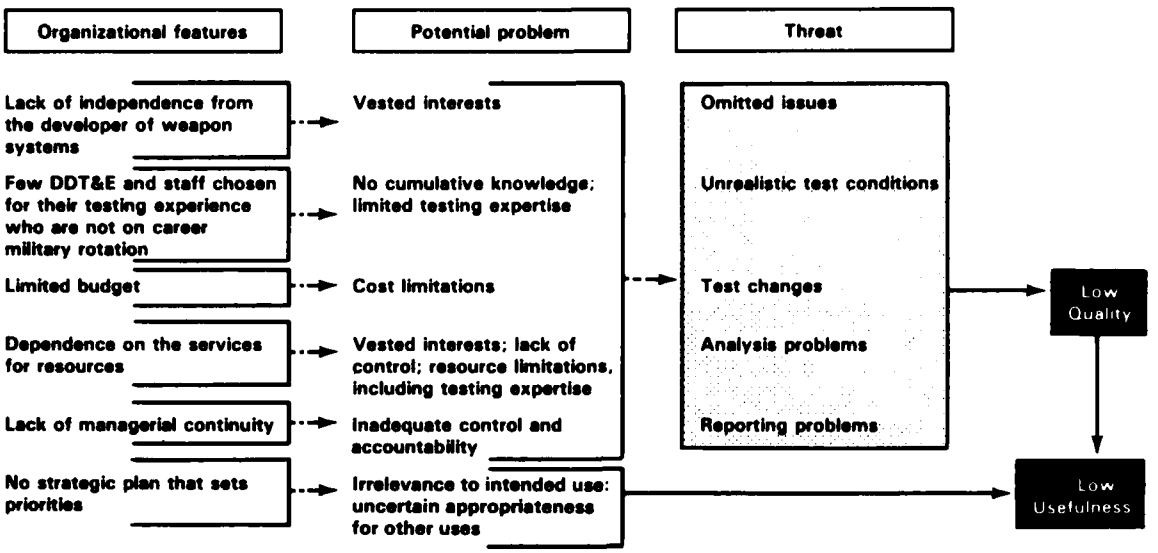
How are JT&E results used? A test's results may be used by its requestor, for the original purpose or for some other, and they may be used by other people. Test results that are not high in quality, are not relevant, do not stipulate the test's limitations, are not timely, and are not presented completely, clearly, and concisely may not be useful for what they were intended for. They may also be used inappropriately.

The requestors made little use of the three tests we examined. The Congress, rather than the DSARC, the requestor, used the IIR Maverick test results--as a reason for denying the Air Force funds for producing the missile system. The Air Force and the DDT&E, rather than DOD's program analysis and evaluation office, the requestor, used ACEVAL's results--for computer models to simulate air combat under different conditions and with missiles with different capabilities. The ACEVAL results have also been cited on both sides of the debate about whether the U.S. weapons acquisition strategy should emphasize quality or quantity. We question the appropriateness of both uses because of the low quality of the test results. The TASVAL results were published after the requester, the Secretary of Defense, left office, so there was no opportunity for him to use them.

The three JT&E's did provide useful information for developing tactics and insights about testing. For example, the test of the IIR Maverick that was conducted in Europe might not have been completed with its quick response to the congressional questions had not the California IIR Maverick test plan and implementation experiences been available. ACEVAL influenced the AMRAAM operational test, in which attempts were made to overcome some of ACEVAL's limitations. The problems in TASVAL, especially with instrumentation, were considered in planning two other JT&E's.

If the quality and usefulness of joint tests and evaluations are flawed, what are the possible reasons? We believe that decisions and events occurring in the test process sometimes lead to joint test results that are unacceptably low in quality and in usefulness. Figure 51 summarizes what we believe the three case studies demonstrate--that the most important threats to the quality of JT&E results and, therefore, to their usefulness are (1) test formulations that fail to consider critical issues, (2) test designs that set up unrealistic test conditions, (3) test implementation that deviates from the test design, (4) test analysis that fails to employ appropriate techniques or to control for

Figure 51
Summary of the Possible Organization-Based Threats to the JT&E Process That Can Lead to Low Quality and Low Usefulness



validity, and (5) test reports that are untimely, based on faulty interpretations, or not appropriately balanced or qualified.

It is true that no test can ever be perfect--for example, not all critical issues can be addressed, realistic test sites are not always available, and implementation must sometimes deviate from what has been planned. However, if there is little understanding of these five basic threats to test quality and of the effect of not addressing them or the need to account for the way they were addressed, then it is almost a foregone conclusion that problems of quality will emerge in the test results. We cannot say with certainty what the reasons are for these threats in the tests we analyzed. The figure shows some possible sources of the problems we observed, some of which may have their origin in the organizational features of JT&E that we have discussed.

The JT&E function belongs to OSD's organization for the development of weapons. Interests can conflict when a developer's weapon systems are called into question by JT&E results. Such conflict clearly would threaten the quality of the results. The responsibility for initiating JT&E's belongs to the DDT&E, whose interests might prevent certain issues from being considered in testing. Neither could happen if the JT&E function were independent of the weapons developer.

The number of members on the DDT&E staff is small, they are not chosen for their testing experience, and they are on career

military rotations. This means that the organization that is responsible for JT&E has no institutional memory and no special expertise in JT&E. It is difficult for the DDT&E to make sure that JT&E's use accumulated experience and appropriate statistical techniques, factors that are critical to the quality of JT&E. This might not be a concern if the JT&E function were in an organization with its own military experts and its own test designers and analysts who could give continuity to the JT&E program and control its quality through designs and implementation. Cumulative knowledge might make it possible to simulate likely combat situations and enemy systems and to bring the various models of weapon performance more into line with what current weapon systems can do.

The JT&E budget is limited, constraining the quality of testing and making it dependent on the services for personnel, instrumentation, equipment, and other resources. The vested interests that the services have in JT&E could easily influence how JT&E's are designed and conducted. The quality of JT&E results might be higher if this were not so.

The management of JT&E's has no continuity from start to finish, a problem for both control and accountability. Although the DDT&E monitors the progress of tests, the staff are few in number, have limited testing experience, and can give little guidance to the tests. It is possible for joint test directors to be chosen who also have limited testing experience. This means that decisions can easily be made without full appreciation of the consequence of the decisions on the quality of the test results.

These organizational features of the JT&E program may threaten not only its quality but also its usefulness, inasmuch as no strategic plan is set for addressing priorities. We have noted the lack of interest among the groups that have potentially the greatest need for JT&E. More interaction in JT&E between the JCS and the services might help in the development of priorities. This, in turn, would make JT&E more responsive and valuable to the decisionmakers who look to the program for information.

In summary, we believe that an independent JT&E organization is necessary if certain threats to JT&E results are to be avoided. However, this is not to say that organizational independence will insure the quality and usefulness of joint tests. Independence will not automatically provide expertise or coordination between the JCS and the services, nor will it necessarily focus the tests on their users' needs for information. We believe that the testing of joint military operations requires the participation of the JCS and the services. It is also apparent, however, that joint tests that do not fulfill their purposes should not be conducted at all. If joint tests and evaluations are to be conducted, the organization that is responsible for them must accumulate acquired knowledge, possess adequate expertise, control the necessary resources, employ the appropriate procedures, and provide timely information

RECOMMENDATIONS TO THE SECRETARY OF DEFENSE

GAO's finding that only 3 of the 13 JT&E's that were completed between 1972 and 1981 focused on joint operations indicates either that DOD does not perceive a need for JT&E information in making decisions about the combinations and structures of forces and the roles and missions of the services or else that DOD does perceive a need for JT&E data in addressing these issues and the JT&E program has not been responsive to this need. GAO recommends that the Secretary of Defense ascertain DOD's need for joint tests that focus on the joint operations of the armed services. The JT&E program should be continued if the Secretary concludes that DOD has such a need.

If the Secretary of Defense determines that DOD does need the JT&E program, GAO recommends that the Secretary take the further steps that are necessary to (1) insure that priorities be established for conducting JT&E's, (2) endow the JT&E program with enough independence, permanence of expert staff, and control of resources to allow the program to conduct and report on joint tests and evaluations that both are high in quality and provide relevant information to their requestors and other users, and (3) require the JT&E program director to develop routine procedures that will insure that thorough records of test data, test results, and their use are maintained.

With regard to the implementation of these recommendations, GAO believes that the recently enacted legislation establishing an office of Operational Test and Evaluation in DOD may provide an opportunity to reduce the problems of JT&E's quality and usefulness that are shown in this report. If JT&E were to become a part of this unit--which, under the legislation, is to be independent of other DOD offices and agencies--then the organizational placement of the JT&E function might no longer pose a potential threat to test quality. However, JT&E's organizational independence is only a necessary condition; it is not in and of itself sufficient for achieving quality and usefulness, because it cannot automatically provide expertise, resources, user focus, or the coordination that is needed between service operations people and test analysts if JT&E's are to be sound.

APPENDIX I

DAVID PRYOR
ARKANSAS

248 Russell Senate Office Building
Washington, D.C. 20510
(202) 224-2353

Arkansas Office:
3000 Federal Building
Little Rock, Arkansas 72201
(501) 378-6336

United States Senate
WASHINGTON, D.C. 20510

COMMITTEES:
AGRICULTURE, NUTRITION, AND FORESTRY
GOVERNMENTAL AFFAIRS
SPECIAL COMMITTEE ON AGING
SELECT COMMITTEE ON ETHICS

May 17, 1982

Mr. Charles Bowsher
Comptroller General
General Accounting Office
Room 7000-A
441 G Street
Washington, D.C. 20548

Dear Mr. Bowsher:

During the past year, I have become increasingly concerned about the testing and evaluation of the weapons and equipment that are being made available to U.S. military personnel in the field. Since we began the present $1.6 trillion five-year defense buildup, the Congress and the Defense Department have been asking questions about how much defense we want but neglecting to address how well our systems will work. If we are truly to commit ourselves to this buildup, we must follow the development and operational testing of the new systems closely so we can be sure that the weapons our soldiers use in the field are effective.

Therefore, I am requesting the General Accounting Office to review several joint operational tests and test evaluations for which certain issues transcend individual service lines and to perform this review in a way that exemplifies the quality, the limitations, and the use of test and evaluation.

Specifically, I would like to understand the limitations and constraints on DOD's test design, implementation, analysis, and reporting activities. Particular questions I am asking GAO to answer include the following:

--Who requests tests and for what reasons?
--Are the test problems defined to include critical operational issues and human factors?
--Do test designs and performance generate valid and reliable data about the operational capabilities of weapon systems, their limitations, and their concepts of employment?
--Are the test results that are reported an accurate reflection of the data collected?
--Do the reports of results address the concerns of those who requested the tests?
--How have the results been used?

PREVIOUS PAGE IS BLANK

APPENDIX I

Because of the important effect answers to such questions can have on weapon system procurement, I would appreciate a draft report by October 1982. I understand that GAO's Institute for Program Evaluation has been reviewing the quality of DOD's joint tests and evaluations. It would be helpful, therefore, if responsibility for this review were assigned to that Institute.

If you have any questions regarding this request, please contact Knox Walkup of my staff at 224-2353.

Sincerely,

David Pryor

DP/kw

BIBLIOGRAPHY

A. GENERAL REFERENCES

1. BDM Corp. "Analysis of Joint Test and Evaluation: Concept and Alternatives." Final report to the Director Defense Test and Evaluation, McLean, Va., June 18, 1979.

2. -----. Joint Test and Evaluation Library, 3 vols. Washington, D.C.: Office of the Under Secretary of Defense, Research and Engineering, Director Defense Test and Evaluation, September 1983.

3. Blue Ribbon Defense Panel. Report to the President and the Secretary of Defense on the Department of Defense. N.p.: July 1, 1970.

4. -----. Report to the President and the Secretary of Defense on the Department of Defense, app. N, Staff Report on Joint Chiefs of Staff Decision-Making. N.p.: July 1, 1970.

5. -----. Report to the President and the Secretary of Defense on the Department of Defense, app. F, Staff Report on Operational Test and Evaluation. N.p.: July 1, 1970.

6. Bridges, Roy D., Jr. "Realism in Operational Test and Evaluation for Close Air Support." Research study, Air University, Maxwell Air Force Base, Ala., May 1976.

7. "Congress Set to Improve Weapons Testing: The Pentagon's Top Scientist Will Lose Responsibility for Weapons Tests in an Effort to Eliminate Waste." Science, August 26, 1983, pp. 19-21.

8. DOD (U.S. Department of Defense), Electronic Warfare During Close Air Support Joint Test Force. "Phase II Test Report Annex M: Lessons Learned." Report from the Joint Test Director, Nellis Air Force Base, Nev., April 7, 1982.

9. -----, Research and Engineering. "Budget Summary: Director of Test and Evaluation, Defense Appropriation." Mimeo, n.p., January 21, 1983.

10. -----. "Initiation of Joint Tests." Memorandum from the Under Secretary, Washington, D.C., March 21, 1983.

11. -----. "Joint Service Test and Evaluation." Paper issued by the Deputy Director Defense Test and Evaluation, Washington D.C., June 1, 1980.

12. -----. Joint Test and Evaluation Procedures Manual. Washington, D.C.: Office of the Secretary of Defense, Director Defense Test and Evaluation, September 1980.

(A. General References, cont'd)

13. DOD, Research and Engineering. "Test and Evaluation." Directive 5000.3, Washington, D.C., December 26, 1979.

14. GAO (U.S. General Accounting Office). Adverse Effects of Large Scale Production of Major Weapons Before Completion of Development and Testing (B-163058). Washington, D.C.: November 19, 1970.

15. -----. Army Operational Test and Evaluation Needs Improvement, CONFIDENTIAL (C-PSAD-80-2). Washington, D.C.: November 13, 1979.

16. -----. Better Planning and Management of Threat Simulators and Aerial Targets Is Crucial to Effective Weapon Systems Performance (MASAD-83-27). Washington, D.C.: June 23, 1983.

17. -----. Development Test and Evaluation of Six Systems (PSAD-79-86). Washington, D.C.: June 25, 1979.

18. -----. DOD Information Provided to the Congress on Major Weapon Systems Could Be More Complete and Useful, SECRET (C-PSAD-80-24). Washington, D.C.: May 9, 1980.

19. -----. Does the Department of Defense Have More Test Capacity Than It Needs? (PSAD-76-75). Washington, D.C.: March 1, 1976.

20. -----. Effectiveness of Testing of Selected Major Weapon Systems, SECRET (PSAD-75-74). Washington, D.C.: June 4, 1975.

21. -----. Follow-on-Operational-Test-and-Evaluation (PSAD-79-1). Washington, D.C.: October 19, 1978.

22. -----. Improvements Needed in Development Testing (B-163058). Washington, D.C.: March 7, 1974.

23. -----. Navy Operational Test and Evaluation: A Valuable Tool Not Fully Utilized, CONFIDENTIAL (PSAD-78-77). Washington, D.C.: March 29, 1978.

24. -----. Need for More Accurate Weapon System Test Results to Be Reported to the Congress, SECRET (PSAD-79-46). Washington, D.C.: March 8, 1979.

25. -----. Operational Test and Evaluation of Foreign Built Systems (PSAD-78-131). Washington, D.C.: July 25, 1978.

26. -----. Operational Testing of Air Force Systems Requires Several Improvements (PSAD-78-102). Washington, D.C.: June 2, 1978.

27. -----. Operational Testing on the Major Caliber Lightweight Gun, CONFIDENTIAL (PSAD-77-4). Washington, D.C.: November 5, 1976.

28. -----. Review of Testing and Evaluation Policies and Procedures (B-163058). Washington, D.C.: April 18, 1974.

29. -----. Review of the Adequacy of Department of Defense Test Resources, CONFIDENTIAL (PSAD-75-84). Washington, D.C.: April 30, 1975.

30. -----. The Importance of Testing and Evaluation in the Acquisition Process for Major Weapon Systems (B-163058). Washington, D.C.: August 7, 1972.

31. -----. Use of the Design for Testability Concept in the Development and Acquisition of Major Weapon Systems (GAO/MASAD-82-38). Washington, D.C.: August 6, 1982.

32. Gordon, Michael R. "Billion-Dollar 'Failures' May Slip Through Pentagon Weapons Testing Net." National Journal, July 24, 1982, pp. 1284-91.

33. Hearings Before the Defense Subcommittee of the House Committee on Appropriations, 98th Cong., 1st sess. (March 16, 1983) (statement of Isham Linder).

34. Waller, Sylvia L. "Operational Test and Evaluation in Support of Executive Decisions." Astronautics and Aeronautics, October 1977, pp. 24-55.

35. "Weapons Testing." Congressional Record (97th Cong., 2nd sess.), April 1, 1982, p. S3322.

B. ACEVAL

1. Amlie, Thomas S. "A Nonstatistical Look at AIMVAL/ACEVAL," SECRET. Report for the Principal Deputy Under Secretary of Defense, Research and Engineering, Defense Advanced Research Projects Agency, Arlington, Va., February 3, 1981.

2. Brown, G. L., et al. Air Combat Effectiveness Study, CONFIDENTIAL. Arlington, Va.: Institute for Defense Analyses, April 1980.

3. Despain, A., et al. Analysis of Relevance to the ACEVAL Experiments, SECRET. Arlington, Va.: SRI International, November 1980.

4. DOD (U.S. Department of Defense), ACEVAL-AIMVAL Joint Test Force. "ACEVAL-AIMVAL," SECRET. Final report briefing, Washington, D.C., date unknown.

(B. ACEVAL, cont'd)

5. DOD, ACEVAL-AIMVAL Joint Test Force. ACEVAL-AIMVAL Final Test Plan, vol. 2, ACEVAL. Nellis Air Force Base, Nev.: August 31, 1976.

6. -----. ACEVAL-AIMVAL Test Plan, vol. 10, Missile Simulation/ Validation Report, CONFIDENTIAL. Nellis Air Force Base, Nev.: August 31, 1977.

7. -----. Air Combat Evaluation Test: Management Lessons Learned. Nellis Air Force Base, Nev.: April 1978.

8. -----, Joint Technical Coordinating Group for Munitions Effectiveness. An Introduction to Air Combat Maneuvering, CONFIDENTIAL. Washington, D.C.: August 1, 1977.

9. -----, U.S. Air Force Tactical Fighter Weapons Center. Air Combat Evaluation (ACEVAL) Follow-on Analysis, SECRET. Nellis Air Force Base, Nev.: October 1978.

10. -----, Weapons Systems Evaluation Group. Multiple Air Combat Evaluation. Arlington, Va.: September 1974.

11. Everson, David. ACEVAL, Final Report, SECRET. Langley Air Force Base, Va.: Tactical Air Command, May 1978.

12. Fay, Robert H., Robert P. McKenzie, and James R. Hildreth. Air Combat Evaluation (ACEVAL), vol. 2, Aircrew Report, SECRET. Nellis Air Force Base, Nev.: ACEVAL-AIMVAL Joint Test Force, February 1978.

13. -----. Air Combat Evaluation (ACEVAL), vol. 1, Executive Summary, SECRET. Nellis Air Force Base, Nev.: ACEVAL-AIMVAL Joint Test Force, February 1978.

14. -----. Air Combat Evaluation (ACEVAL), vol. 3, Test Data, SECRET. Nellis Air Force Base, Nev.: ACEVAL-AIMVAL Joint Test Force, February 1978.

15. -----. Air Combat Evaluation (ACEVAL), vol. 4, Test Operations and Analysis, SECRET. Nellis Air Force Base, Nev.: ACEVAL-AIMVAL Joint Test Force, February 1978.

16. Higginbotham, K. F., et al. Multiple Air Combat Effectiveness (MACE) Model, CONFIDENTIAL. Arlington, Va.: Institute for Defense Analyses, October 1979.

17. Kaufman, I. A. Multiple Air Combat Evaluation (ACEVAL), SECRET. Arlington, Va.: Institute for Defense Analyses, January 1979.

18. -----. "Multiple Air Combat Evaluation (ACEVAL)," SECRET. Extracts from IDA Log Nos. HQ 78-20666/1, HQ 78-20655/1, and

HQ 78-20656/1, Institute for Defense Analyses, Arlington, Va., August 5, 1981.

19. -----, et al. Design of an Operational Test for the Evaluation of Multiple Air-to-Air Combat (ACEVAL), CONFIDENTIAL. Arlington, Va.: Institute for Defense Analyses, December 1975.

20. -----. Multiple Air Combat Evaluation. Arlington, Va.: Institute for Defense Analyses, September 1974.

21. Kelly, C. T., Jr., H. G. Hoover, and M. D. Miller. Acquisition Options for Fighter Aircraft: The ACEVAL Test, CONFIDENTIAL. Santa Monica, Calif.: Rand, November 1980.

22. Kross, Walt. "ACEVAL/AIMVAL: Abusing Atari in the Desert." Armed Forces Journal International, January 1982, pp. 52-56.

23. Powell, Cecil W. "ACEVAL: A Critical Analysis of a Proposed Air Combat Evaluation Program," CONFIDENTIAL. Research report, Air War College, Air University, Maxwell Air Force Base, Ala., April 1975.

24. Schemmer, Benjamin F. "The One-and-a-Half-Trillion-Dollar Misunderstanding." Washingtonian, August 1981, pp. 3-4.

25. "The Department of Defense Program of Test and Evaluation, FY 1975." Hearings Before the Defense Subcommittee of the Senate Committee on Appropriations, 93rd Cong., 2nd sess. (April 5, 1974) (statement of Alfred D. Starbird).

26. Tissot, Ernest E., and James R. Hildreth. Air Intercept Missile Evaluation (AIMVAL) Final Report, vol. 1, Summary, SECRET. Nellis Air Force Base, Nev.: ACEVAL-AIMVAL Joint Test Force, September 6, 1977.

27. Veda, Inc. AMRAAM Operational Utility Evaluation: Final Report, SECRET. Las Vegas, Nev.: May 4, 1981.

28. Wood, Rodney J., Peter A. Borgo, and Roger K. Hoppe. High Energy Laser Systems ACEVAL/AIMVAL Data Analysis, SECRET. Albuquerque, N. Mex.: BDM Corp., October 1979.

C. IIR MAVERICK

1. DOD (U.S. Department of Defense), Office of the Deputy Secretary. "DSARC II Decision on Imaging Infrared Maverick." Memorandum for the secretaries of the military departments, Washington, D.C., November 19, 1976.

2. -----, Office of the Secretary. "AGM-65D Imaging Infrared Maverick." Memorandum for the Secretary of the Air Force, Washington, D.C., May 6, 1983.

(C. IIR MAVERICK, cont'd)

3. DOD, Office of the Secretary. "Imaging Infrared (IIR) Maverick." Memorandum for the Secretary of the Air Force, Washington, D.C., September 29, 1982.

4. -----. "Imaging Infrared (IIR) Maverick Decision Memorandum." Memorandum for the Secretary of the Air Force, Washington, D.C., March 29, 1982.

5. -----, Research and Engineering. "Imaging Infrared Maverick Joint Operational Test and Evaluation." Memorandum for secretaries of the military departments from the Office of the Director, Washington, D.C., November 26, 1976.

6. -----, U.S. Air Force. Air Force Manual, 2-1, Aerospace Operational Doctrine: Tactical Air Operations--Counter Air, Close Air Support, and Air Interdiction. Washington, D.C.: May 2, 1969.

7. -----. Decision Coordinating Paper No. 154. Washington, D.C.: August 23, 1976.

8. -----. "Selected Acquisition Report: IIR Maverick AGM-65D as of Date 31 December 1976," CONFIDENTIAL. Washington, D.C.: Office of the Secretary of Defense, January 5, 1977.

9. -----. Test and Evaluation Master Plan (TEMP) for Imaging Infrared Maverick. Washington, D.C.: April 15, 1980.

10. -----, U.S. Air Force, Office of the Assistant Chief of Staff, Studies and Analyses. Saber Test (Bravo): Summary Report on the AF/SA Evaluation of IIR Maverick JOT&E, SECRET. Washington, D.C.: March 16, 1977.

11. -----, U.S. Air Force Test and Evaluation Center. Air Operations Plan: IIR Maverick Joint Operational Test and Evaluation. Langley Air Force Base, Va.: Tactical Air Command, February 1977.

12. -----. Data Management Plan: IIR Maverick Joint Operational Test and Evaluation. Kirtland Air Force Base, N. Mex.: February 1977.

13. -----. IIR Maverick Joint Operational Test and Evaluation Plan. Kirtland Air Force Base, N. Mex.: January 1977.

14. Flanagan, T. Patrick. AGM-65D IR Maverick Initial Operational Test and Evaluation Plan. Kirtland Air Force Base, N. Mex.: U.S. Air Force Test and Evaluation Center, October 1, 1980.

15. Freathy, Alfred L., and Jessie K. Beavers. AGM-65D Infrared Maverick Initial Operational Test and Evaluation Final

Report, SECRET. Kirtland Air Force Base, N. Mex.: U.S. Air Force Test and Evaluation Center, December 1982.

16. GAO (U.S. General Accounting Office). Critical IR Maverick Issues Remain Unresolved After Five Years of Operational Testing, SECRET (GAO/C-IPE-82-1). Washington, D.C.: June 25, 1982.

17. -----. Letter to The Honorable David Pryor, U.S. Senate, and statement of fact entitled "The USAF's IR Maverick Program," Washington, D.C., May 4, 1983.

18. Latta, L. D. Imaging Infrared Guidance Unit Countermeasures Static Tests, SECRET. White Sands Missile Range, N. Mex.: Joint Services Electro-optical Guided Weapons Countermeasures Test Program, July 7, 1975.

19. -----, and Ernesto Ruiz. Countermeasures Evaluation of the IIR Maverick JOT&E, SECRET. White Sands Missile Range, N. Mex.: Joint Services Electro-optical Guided Weapons Countermeasures Test Program, November 23, 1977.

20. Macleay, Lachlan. Program Management Plan for AGM-65 Missile Program, SECRET. Wright-Patterson Air Force Base, Ohio: U.S. Air Force Aeronautical Systems Division, May 1975.

21. Madison, James A., and Augustine R. Letto. Imaging Infrared (IIR) Maverick Joint Operational Test and Evaluation, CONFIDENTIAL. Kirtland Air Force Base, N. Mex.: U.S. Air Force Test and Evaluation Center, July 1977.

22. Martin, James R., and Michael T. Reich. Imaging Infrared (IIR) Tracker European Test and Evaluation, CONFIDENTIAL. Kirtland Air Force Base, N. Mex.: U.S. Air Force Test and Evaluation Center, September 1978.

23. Stahl, F. G., et al. Joint Operational Test and Evaluation of the Imaging Infrared Maverick System, SECRET. Arlington, Va.: System Planning Corp., August 1, 1977.

24. Stahl, F. G., M. H. Crowell, and L. M. Silvester. Testing of the Imaging Infrared Maverick System: Issues and Analyses, SECRET. Arlington, Va.: System Planning Corp., July 1979.

25. U.S. House of Representatives. Hearings Before the Committee on Armed Services, Subcommittee on Research and Development. USAF Development of Imaging Infrared Maverick Seeker Tracker. 96th Cong., 1st sess. (April 17, 1978).

26. -----, Committee on Armed Services. Letter from the Chairman, Research and Development Subcommittee, to the Secretary of the Air Force, Washington, D.C., August 12, 1977.

D. TASVAL

1. Bahnsen, John C., and David R. Brown. "Final Report of the Joint Attack Weapons System Tactics Development and Evaluation," SECRET. Working paper, U.S. Army Training and Doctrine Command, Washington, D.C., November 1977.

2. Bustillos, Ruben, and Servando Hernandez. Army TASVAL Independent Evaluation Plan. Alexandria, Va.: Defense Logistics Agency, June 1979.

3. DOD (U.S. Department of Defense), Research and Engineering. "Joint Test and Evaluation of Tactical Aircraft Effectiveness and Survivability in Close Air Support Anti-Armor Operations." Memorandum for the secretaries of the Army, Navy, and Air Force, Washington, D.C. September 19, 1977.

4. -----. "TASVAL," SECRET. Memorandum for the record from Office of the Under Secretary of Defense, Director Defense Test and Evaluation, Washington, D.C., July 26, 1981.

5. -----. "TASVAL Joint Test Objectives." Memorandum for the assistant secretaries of the Army, Navy, and Air Force, the TASVAL Joint Test Director, and the Institute for Defense Analyses from Office of the Under Secretary of Defense, Director Defense Test and Evaluation, Washington, D.C., August 23, 1979.

6. -----, TASVAL Joint Test Force. Joint Test of Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL), vol. 1, SECRET. Washington, D.C.: Office of the Under Secretary of Defense, Research and Engineering, Director Defense Test and Evaluation, May 1980.

7. -----. Joint Test of Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL), vol. 2, SECRET. Washington, D.C.: Office of the Under Secretary of Defense, Research and Engineering, Director Defense Test and Evaluation, May 1980.

8. -----, U.S. Air Force, Assistant Chief of Staff, Studies and Analyses. Saber Test Echo: An Analysis of the TASVAL Joint Test Tactical Aircraft Effectiveness and Survivability in CAS Anti-Armor Operations, SECRET. Washington, D.C.: June 1980.

9. -----, U.S. Army. Briefing on analysis of TASVAL data by U.S. Army Training and Doctrine Command, Systems Analysis Activity, February 3, 1982.

10. -----. Briefing on Nellis Air Force Base, Nev., March 19-20, 1981, meeting on joint Army-Air Force resolution of TASVAL critical issues and recommendations, February 3, 1982.

11. -----. Briefing on scout and attack helicopter training and teamwork needs learned from TASVAL, February 3, 1982.

12. -----, U.S. Army Aviation Center. Joint Air Attack Team Operations, TAC 50-20, TRADOC 17-50-3. Ft. Rucker, Ala.: April 30, 1979.

13. -----, U.S. Army Combat Development Experimentation Command. Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL), vol. 1, Test Plan. Ft. Ord, Calif.: March 1979.

14. -----. Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL): Lessons Learned. Ft. Ord, Calif.: December 1979.

15. -----, U.S. Army Training and Doctrine Command. Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL), vol. 3, Appendix I, CONFIDENTIAL. Washington, D.C.: April 1981.

16. -----. Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL), vol. 4, Appendixes J and K, CONFIDENTIAL. Washington, D.C.: April 1981.

17. -----. Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL), vol. 5, Appendixes L and M, CONFIDENTIAL. Washington, D.C.: April 1981.

18. -----. Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL), vol. 6, Appendix N, CONFIDENTIAL. Washington, D.C.: April 1981.

19. -----. Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL), vol. 7, Appendix O, CONFIDENTIAL. Washington, D.C.: April 1981.

20. -----. Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL), vol. 8, Appendix P, CONFIDENTIAL. Washington, D.C.: January 1981.

21. -----. Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL), vol. 1, Executive Summary, CONFIDENTIAL. Washington, D.C.: Janury 1981.

22. -----. Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL), vol. 2, Main Report and Appendixes A-H, CONFIDENTIAL. Washington, D.C.: April 1981.

(D. TASVAL, cont'd)

23. Freeh, James L., John N. Donis, and Robert L. McDaniel. Analysis of AAH Issues: Final Report, SECRET. Arlington, Va.: System Planning Corp., November 1979.

24. GAO (U.S. General Accounting Office). Report on aerial fire support programs, OSD Case 4922. Letter PSAD-78-119 to the Secretary of Defense, Washington, D.C., June 6, 1978.

25. Leatherbury, C. H., et al. Joint Test and Evaluation of Tactical Aircraft Survivability: Test Design, vol. 2, Appendix B: Preliminary Appraisal of Problems Arising from Use of U.S. or NATO Surrogates for Soviet Air Defense Systems, SECRET. Arlington, Va.: Institute for Defense Analyses, December 1977.

26. -----. Joint Test and Evaluation of Tactical Aircraft Survivability: Test Design, vol. 1, Main Paper, CONFIDENTIAL. Arlington, Va.: Institute for Defense Analyses, December 1977.

27. -----. "Operational Test and Evaluation of Tactical Aircraft Survivability in Anti-Armor Operations (TASVAL)," 2 vols., SECRET. Draft report, Institute for Defense Analyses, Arlington, Va., December 8-11, 1980.

28. -----. Operational Test and Evaluation of Tactical Aircraft Survivability in Anti-Armor Operations (TASVAL), SECRET. Arlington, Va.: Institute for Defense Analyses, December 1980.

29. McGinty, Kenneth R., and Roth S. Schleck, Jr. "Tests: A Mean to Train and Learn." TASVAL Review, June 1980, pp. 10-13.

30. Neuwien, R. A., et al. TD&E Joint Countering Attack Helicopters (J-CATCH) Report, Phases I-II, SECRET. Ft. Monroe, Va.: U.S. Army Training and Doctrine Command, November 19, 1979.

31. Porter, R. A., and S. J. Hubbard. Validation of Close Air Support (CAS) Phase II Results, Including IDA Study S-472, SECRET. Arlington, Va.: Institute for Defense Analyses, February 1976.

32. Prouty, James Russell. "An SA-8 Versus A-10 Engagement Model Developed from TASVAL Results," SECRET. Thesis, Naval Postgraduate School, Monterey, Calif., September 1980.

33. Tice, Jim. "AF, Army Pilots Hone Antitank Skills." Air Force Times, April 26, 1982, p. 34.

34. U.S. Senate. Hearings Before a Subcommittee of the Committee on Appropriations. Department of Defense Appropriations for Fiscal Year 1979. 95th Cong., 2nd sess. (1978).

SUPPLEMENTARY DATA ON DOD'S JT&E ACTIVITY 1972-83

This appendix contains the following items: (1) a reprint of Public Law 98-94, title XII, part B, section 1211, establishing the position of Director of Operational Test and Evaluation on November 1, 1983; (2) a reprint of the November 22, 1982, memorandum in which the Deputy Director for Defense Test and Evaluation requested the participation of the Joint Chiefs of Staff in the JT&E program; (3) a reprint of the January 22, 1981, memorandum in which the Navy withheld its concurrence with the 1980 JT&E procedures manual; (4) a list of the 30 JT&E's initiated between 1972 and 1983 with summary data on their requestors, objectives, duration, and cost.

97 STAT. 684 PUBLIC LAW 98-94—SEPT. 24, 1983

[Title XII--General Provisions]

. . . . .

PART B—DEPARTMENT OF DEFENSE MANAGEMENT MATTERS

ESTABLISHMENT OF DEFENSE DIRECTOR OF OPERATIONAL TEST AND EVALUATION

SEC. 1211. (a)(1) Chapter 4 of title 10, United States Code, is amended by inserting after section 136 the following new section:

10 USC 136a.

"§136a. Director of Operational Test and Evaluation: appointment; powers and duties

"(a)(1) There is a Director of Operational Test and Evaluation in the Department of Defense, appointed from civilian life by the President, by and with the advice and consent of the Senate. The Director shall be appointed without regard to political affiliation and solely on the basis of fitness to perform the duties of the office of Director. The Director may be removed from office by the President. The President shall communicate the reasons for any such removal to both Houses of Congress.

Definitions.

"(2) In this section:

"(A) 'Operational test and evaluation' means—

"(i) the field test, under realistic combat conditions, of any item of (or key component of) weapons, equipment, or munitions for the purpose of determining the effectiveness and suitability of the weapons, equipment, or munitions for use in combat by typical military users; and

"(ii) the evaluation of the results of such test.

"(B) 'Major defense acquisition program' means a Department of Defense acquisition program that is a major defense acquisition program for purposes of section 139a(a)(1) of this title or that is designated as such a program by the Director for purposes of this section.

10 USC 139a.

"(b) The Director is the principal adviser to the Secretary of Defense on operational test and evaluation in the Department of Defense and the principal operational test and evaluation official within the senior management of the Department of Defense. The Director shall—

"(1) prescribe, by authority of the Secretary of Defense, policies and procedures for the conduct of operational test and evaluation in the Department of Defense;

"(2) provide guidance to and consult with the Secretary of Defense and the Secretaries of the military departments with respect to operational test and evaluation in the Department of Defense in general and with respect to specific operational test and evaluation to be conducted in connection with a major defense acquisition program;

"(3) monitor and review all operational test and evaluation in the Department of Defense;

"(4) coordinate operational testing conducted jointly by more than one military department or defense agency;

"(5) analyze the results of the operational test and evaluation conducted for each major defense acquisition program and, at the conclusion of such operational test and evaluation, report to the Secretary of Defense and to the Committees on Armed

Services and on Appropriations of the Senate and House of Representatives as provided in subsection (c) on—

"(A) whether the test and evaluation performed was adequate; and

"(B) whether the test and evaluation results confirm that the items or components actually tested are effective and suitable for combat; and

"(6) review and make recommendations to the Secretary of Defense on all budgetary and financial matters relating to operational test and evaluation, including operational test facilities and equipment, in the Department of Defense.

"(c) Each report of the Director required under subsection (b)(5) shall be submitted to the committees specified in that subsection in precisely the same form and with precisely the same content as the report originally was submitted to the Secretary and shall be accompanied by such comments as the Secretary of Defense may wish to make on such report. Report submittal.

"(d) The Director reports directly, without intervening review or approval, to the Secretary of Defense. The Director shall consult closely with, but the Director and the Director's staff are independent of, the Under Secretary of Defense for Research and Engineering and all other officers and entities of the Department of Defense responsible for research and development.

"(e)(1) The Secretary of a military department shall report promptly to the Director the results of all operational test and evaluation conducted by the military department and of all studies conducted by the military department in connection with operational test and evaluation in the military department.

"(2) The Director may require that such observers as he designates be present during the preparation for and the conduct of the test part of any operational test and evaluation conducted in the Department of Defense.

"(3) The Director shall have access to all records and data in the Department of Defense (including the records and data of each military department) that the Director considers necessary to review in order to carry out his duties under this section. Records and data accessibility.

"(f)(1) Operational testing of a major defense acquisition program may not be conducted until the Director has approved in writing the adequacy of the plans (including the adequacy of projected levels of funding) for operational test and evaluation to be conducted in connection with that program.

"(2) A final decision within the Department of Defense to proceed with a major defense acquisition program beyond low-rate initial production may not be made until the Director has submitted to the Secretary of Defense the report with respect to that program required by subsection (b)(5) and the Committees on Armed Services and on Appropriations of the Senate and House of Representatives have received that report.

"(g)(1) The Director shall prepare an annual report summarizing the operational test and evaluation activities of the Department of Defense during the preceding fiscal year. Each such report shall be submitted concurrently to the Secretary of Defense and the Congress not later than January 15 immediately following the end of the fiscal year for which the report is prepared. The report shall include such comments and recommendations as the Director considers appropriate, including comments and recommendations on resources and facilities available for operational test and evaluation Report submittal.

and levels of funding made available for operational test and evaluation activities. The Secretary may comment on any report of the Director to Congress under this paragraph. Comments.

"(2) The Director shall comply with requests from Congress (or any committee of either House of Congress) for information relating to operational test and evaluation in the Department of Defense.

"(h) The President shall include in the Budget transmitted to Congress pursuant to section 1105 of title 31 for each fiscal year a separate statement of estimated expenditures and proposed appropriations for that fiscal year for the activities of the Director of Operational Test and Evaluation in carrying out the duties and responsibilities of the Director under this section.". 96 Stat. 908.

(2) The table of sections at the beginning of such chapter is amended by inserting after the item relating to section 136 the following new item:

"136a. Director of Operational Test and Evaluation: appointment; powers and duties.".

(b) Section 5315 of title 5, United States Code, is amended by adding at the end thereof the following new item:

"Director of Operational Test and Evaluation, Department of Defense.".

(c) The amendments made by this section shall take effect on November 1, 1983. Effective date. 10 USC 136a note.

APPENDIX III

MEMORANDUM DATED NOVEMBER 22, 1982

The following memorandum on joint test and evaluation was sent from the Deputy Director for Defense Test and Evaluation to the Director of the Joint Chiefs of Staff:

"The Director Defense Test and Evaluation (DDT&E) has the responsibility to administer the Joint Test and Evaluation (JT&E) Program. To have a successful program it is important, in my view, to have a coherent strategy behind the proposals for joint tests, especially with the varied and sometimes divergent views of the Services. This coherency, I feel, can be enhanced by active participation by the JCS.

"The primary purpose of JT&E is to examine the capability of developmental and deployed systems to perform their intended missions in a joint environment. JT&E's may also be conducted to provide information in the following areas: technical concepts evaluation; systems requirements; system improvements; system interoperability; force structure planning; testing methodologies; and doctrine, tactics, and operational procedures for joint operations. As you can see there is a wide ranging scope of possibilities.

"There are currently six ongoing joint tests. We are now considering possible selections for a test or tests to begin in FY 85 from four proposals. Those under consideration are: Target Engagements Using Labor Designators; Joint Chemical Warfare; Air to Air Missile Concept Evaluation and Joint Attack of Deep Targets. Briefings on these candidates are available to you or your staff if desired. These efforts were recently briefed to a panel which I head, and on which the JCS was represented.

"I believe the joint test program can be very beneficial in the joint arena, and we must insure that those tests selected provide answers to our most important questions. However, I am not in a position to measure our selections. I feel that we are examining worthwhile candidates, but their relative worth is not apparent. I solicit your help in the review of joint test nominations to help us insure that we are best serving DoD's interests."

MEMORANDUM DATED JANUARY 22, 1981

The following memorandum on "Joint Test and Evaluation Procedures Manual" was sent from the Acting Assistant Secretary of the Navy for Research, Engineering, and Systems to the Principal Deputy Under Secretary of Defense for Research and Engineering. "Reference (a)" is to "DOD Directives System Coordination and Control Record of 19 Sep 1980":

"The purpose of this memorandum is to advise you that the Navy non-concurs with promulgation of the Joint Test and

Evaluation Procedures Manual which was proposed by reference (a).

"Our concern is broadly based, involving the fundamental structure and procedures of the DoD JT&E Program. The current process essentially 'fences' a portion of the Defense Budget each year for JT&E, then searches out candidates for joint testing, and finally 'taxes' the Services to provide the necessary personnel, weapons systems, and support funding. This procedure would not be unreasonable if high-priority candidates for JT&E abounded, and if the results of past efforts demonstrably justified the money and services expended of [sic] them. In the Navy's view, neither is self evident. We can identify few JT&E candidates; and our assessment of the value of a decade's JT&E projects is not confidence-inspiring. JT&E is highly appropriate and valuable, it may well be that reserving the JT&E processes for such ad hoc projects would result in better stewardship of the taxpayers' dollars than does the current 'fencing' procedure. Fiscal prudence suggests that in these austere times we should not be hesitant about allowing JT&E projects to compete with alternative uses for Defense funds.

"Accordingly, before further institutionalizing past and present JT&E practices by promulgation of the JT&E Procedures Manual, it would be wise to conduct a careful review of the results these practices have produced from the 'users' point of view. It is notable that the 1979 JT&E analysis sponsored by DDT&E contained no such utility assessment.

"I recommend that issuance of the JT&E Procedures Manual be deferred; and that the Director, Defense Test and Evaluation initiate a DoD-wide users review of JT&E, to weigh the explicit benefits achieved from past testing against the total costs incurred."

THE 30 JT&E'S INITIATED 1972-83

On the facing page and on the four pages following it, we list in chronological order the 30 JT&E's that DOD initiated between 1972 and 1983. We give the title of each test, a summary of its objectives, its duration, its cost to DDT&E, and the name of its requestor. Notes to the table are on page 161.

Test	Objective	FY duration	DDT&E costs (thousands)	Requestor
Electro-optical (EO) Maverick (Combat Hunter)	Determine EO Maverick's ability to attack and destroy armored vehicles in a Soviet combined arms unit in Europe and the cost in attrition of U.S. aircraft by air defense units.	1972-73 a/	$ 3,186	OSD
Aircraft Survivability (VULEVAL, also known as SEAS)	Understand aircraft vulnerability to nonnuclear munitions, such as 16.5-, 23-, and 57-mm projectiles.	1972-75 b/	$10,231	Joint Tactical Coordination Group on Aircraft Survivability
Air-to-Air Weapons Effectiveness (AIRVAL)	Develop and demonstrate improved instruments and analytical methods for predicting the performance of air-to-air missiles; evaluate air-to-air missile systems AIM-7 and AIM-9; improve tactics; provide data for future systems design.	1972-75 a/	$ 5,564	DDR&E
Radar Bombing Accuracy (RABVAL)	Analyze performance of F-111F and A-6E tactical aircraft radar bombing system during simulated combat.	1972-75 a/	$ 5,406	DDT&E
Electronic Warfare (EW, EWJT, or EWARVAL)	Determine relative effectiveness of operational air-to-ground electronic warfare systems supporting tactical strikes against integrated Soviet air defense system.	1972-76 a/	$10,507	DDR&E
Hit Probability (HITVAL)	Validate and improve models used to determine probability of antiaircraft guns hitting fixed- and rotary-wing aircraft during close air support in Europe.	1973-75 a/	$11,130	DDR&E

APPENDIX III

Test	Objective	FY duration	DDT&E costs (thousands)	Requestor
Laser Guided Weapons Countermeasures (LGW/CM)	Determine which countermeasures are effective and practical against U.S. laser weapons.	1973-75 c/	$ 5,776	OSD
Airborne Target Acquisition (SEEKAL)	Evaluate alternative systems and techniques for acquiring ground targets in close air support; determine the effectiveness of two types of target acquisition simulator for such evaluation.	1973-76 a/	$ 6,057	DDR&E in response to Congress
A-7/D/A-10 Flyoff	Determine the relative operational effectiveness of A-7 and A-10 aircraft in close air support in European weather.	1974 a/	$ 1,073	Senate Armed Services Committee
Laser Guided Weapons in Close Air Support (LGW/CAS)	Determine the command and control system's limit on the number of effectively employable laser-guided weapons and related laser designators in close air support.	1974 d/	$ 3,127	OSD
Close Air Support Command and Control (CAS/C^2)	Determine response time, communications requirements, and integration of close air support with other tactical combat operations.	1974-76 a/	$ 3,147	Dep. Sec. DOD in response to Congress
Forward Area Defense (ADVAL or FAD)	Assess the operational effectiveness of forward-area U.S. air defense systems in joint field exercises.	1975 e/	$ 535	DDT&E upon IDA recommendation
Short Range Air-to-Air Missile or Air Intercept Missile Evaluation (AIMVAL)	Determine the contribution of off-bore sight and seekers to short-range air-to-air missiles in combat within visual range.	1975-77 a/	$12,444	DDR&E in response to Congress

APPENDIX III

Test	Objective	FY duration	DDT&E costs (thousands)	Requestor
Logistics-over-the-Shore or Operational Resupply (LOTS or J-LOTS)	Evaluate the ability of the services to deploy LOTS units and handle container ships, barges, and cargo over the shore.	1975-78 a/	$ 6,261	Army
Multiple Air-to-Air Combat (ACEVAL)	Determine how combat between specific aircraft systems is affected by the number of aircraft and other factors; evaluate the test methodology.	1975-78 a/	$ 9,065	Program Analysis and Evaluation, OSD
Ground Target Engagement (TEVAL)	Evaluate aircraft sensors, ground equipment aids, command and control configurations, and other airborne target-engagement systems.	1976 e/	$ 84	Not available
Electronic Warfare During Close Air Support (EW/CAS)	Evaluate electronic countermeasures, defense suppression, and tactics against Soviet jamming of U.S. tactical communications with rotary- and fixed-wing aircraft in close air support.	1976	$63,213 f/	DDT&E
Electro-Optical Guided Weapons Countermeasures (E-O GW CM)	Determine the vulnerability of U.S. electro-optical guided weapons to enemy countermeasures.	1976 to ? g/	Average of $4,725 each FY 1981-83	OSD
Data Link Vulnerability (DVAL)	Develop and validate a simulated and a field test method of assessing the performance of data links (including tactical, control, weapon control, command, control, communications, and reconnaissance) in a hostile electronic environment.	1977	$18,063 f/	DDT&E

APPENDIX III

Test	Objective	FY duration	DDT&E costs (thousands)	Requestor
Imaging Infrared (IIR) Maverick	Examine the operational effectiveness of the IIR Maverick missile with various aircraft and target locator-designator combinations in battle.	1977 a/	$ 1,709	Dep. Sec. DOD
Tactical Aircraft Effectiveness and Survivability in Anti-Armor Operations (TASVAL)	Evaluate the separate and joint ability of rotary- and fixed-wing aicraft to destroy enemy armor and survive in European close air support.	1977-81 a/	$20,743	OSD
Advanced Anti-Armor Combat Vehicle (ARMVAL)	Explore the effectiveness of combined arms forces using lightweight, agile, surrogate vehicles in combat against conventional armor.	1978-84	$13,017 f/	Asst. Sec. for Program Analysis and Evaluation
Identification of Friend, Foe, or Neutral (IFFN)	Evaluate NATO air defense command-and-control ability to discriminate friends, enemies, and neutrals and assess near-term strengthening of procedural and equipment weaknesses.	1978-87	$95,208 f/	DDR&E upon Defense Science Board recommendation
Tube Launched Guided Projectiles (TLGP)	Assess a technologically advanced, antiarmor, fire-and-forget, 105-mm tube-launched nonimaging infrared seeker guided projectile.	1979 d/	0	Not available
Command, Control, and Communications Countermeasures (Counter C^3/CM)	Assess effectiveness, develop tactics, and identify system improvements for U.S. forces' countering Soviet command, control, and communications.	1979-89	$70,465 f/	OSD
Central Region Airspace Control Plan (CRACP)	Evaluate fast-time simulated, central region airspace control systems designed to resolve conflicts among friendly forces in common forward-combat airspace.	1980-82	$ 565	DDT&E

APPENDIX III

Test	Objective	FY duration	DDT&E costs (thousands)	Requestor
Theater Air Defense (TAD)	Improve command and control of friendly short-range air defense assets and reduce air casualties from friendly fire.	1981-82 h/	$ 232	JCS, Air Force
Forward Area Air Defense (JFAAD)	Evaluate joint command and control of airspace during defensive operations.	1981-87	$65,706 f/	DDT&E upon IDA recommendation
Joint Logistics-over-the Shore II (J-LOTS II)	Determine the effectiveness of joint over-the-shore container discharge and movement into a temporary facility at a site with state-three seas.	1981-87	$22,400 f/	Navy
Joint Direction Finding (JDF)	Determine whether services' radio direction finding equipment can be joined to support joint tactical operations.	1982 h/	$ 0	CINCPAC (JCS)

a/One of the 13 completed JT&E's from which we selected 3 as case studies.

b/Sponsorship by DDT&E concluded in 1975; program continued under Tri-Service funding.

c/Expanded to include the full spectrum of electro-optical weapons as E-O GW CM.

d/Terminated in 1979.

e/Terminated in 1977.

f/Projected "costs to completion" in January 21, 1983.

g/This DDT&E-chartered test is not typical of joint tests. It is a continuous, independent program headed by a civilian who reports directly to the DDT&E for test policy, program direction, planning, execution, and reporting. All electro-optical weapons developed by the services are tested by this test group. Its reports are signed by the group's director and the service involved.

h/Terminated in 1982.

APPENDIX III

TECHNICAL DATA FOR CHAPTER 4

ON IIR MAVERICK

The items in this appendix are supplementary and supporting technical data for the presentation of our analysis in chapter 4. The items are keyed by number to the text, where they are cited as parenthetical references to this appendix. They are also listed immediately below, each with its item number and a short descriptive title of its contents.

ITEM		Page
1	Chronology of the IIR Maverick JT&E	162
2	Description of the IIR Maverick missile	163
3	DDT&E's DSARC II statement	163
4	Operational uncertainties listed by the DDT&E	164
5	Deficiencies of planned Air Force tests	164
6	The phases of employment of the IIR Maverick	165
7	Operational uncertainties to be understood	165
8	The JTF's original data analysis plan	166
9	The characteristics of proposed and actual test missions with "good" visibility and "poor"	169
10	A comparison of A-10 target-area acquisition ranges	169
11	Passes with target detection ranges of 500-1,000 feet less than wings-level range	170
12	Number of passes by type of mission	170
13	IIR Maverick test director's comments on countermeasures report	170
14	Percentages of valid target acquisition in four reports	171
15	Mean launch and abort ranges by IIR Maverick mission	172
16	Number of valid and invalid targets with and without Pave Penny in A-10 aircraft in close air support	173
17	Simulated launch ranges	173
18	Tracking time by pass outcome	174
19	Countermeasure results	175
20	The relation between tank temperature and launch range	176
21	Gaps in the thermal data	176

ITEM 1: Chronology of the IIR Maverick JT&E

September 28, 1976	DSARC II reviews IIR Maverick program
October 14, 1976	Deputy Director for Research and Engineering reviews planned Air Force tests
November 19, 1976	DSARC II issues decision memorandum
November 26, 1976	DDT&E issues memorandum for IIR Maverick JT&E
December 10, 1976	Air Force accepts lead role for JT&E and designates joint test director
December 20, 1976	DDT&E approves nominations for joint test director and deputy test directors
January 1977	AFTEC presents test plan
February 1977	Joint test conducted at Ft. Polk, Louisiana
July 1977	JTF publishes report
August 1977	System Planning Corporation publishes report

ITEM 2: Description of the IIR Maverick missile

Procurement for the IIR Maverick program began in April 1974 with the purpose of providing the U.S. Air Force Tactical Air Command with

> "a rocket propelled, air-to-surface precision guided missile that develops tracking signals from the naturally occurring thermal energy of the target. It is designed to destroy small hard tactical targets during day or night even under limited adverse weather conditions in the counter-air, interdiction, and close-air support operations of the tactical air forces." (II.C.8, p. 2)

According to the Air Force, the IIR Maverick missile, the AGM-65D, is intended to be an improvement over the television Maverick because it can be operated at night and with poor visibility.

More specifically in regard to the operation of the missile, when a member of the aircrew points the seeker at the target area, the enhanced target contrast provided in the infrared spectrum permits target acquisition and lock-on at long stand-off ranges with only the cockpit monitor. The infrared seeker senses minute differences in temperature and, with the help of a mechanical scanning system, produces a TV-like display of temperature gradients on the cockpit monitor. The emission of temperature from all objects does not depend on either nature (as from the heat of the sun) or artificial heat (as from flares, spotlights, and the like). This means the IIR Maverick missile system should be operable not only in daylight but also at night and when visibility is poor.

In the March 1982 systems acquisition report, the wording "adverse weather" was changed to "limited adverse weather," without explanation or definition. The Air Force states that the IIR Maverick will be used in counter-air operations, but we find no operational testing of it in a counter-air scenario. Another phase of the Ft. Polk IIR Maverick test was planned, in which three sorties were to be flown against parked aircraft targets, but they never took place.

ITEM 3: DDT&E's DSARC II statement

At the September 1976 DSARC II meeting on the IIR Maverick missile system, the Director for Defense Test and Evaluation made remarks to the following effect. The Air Force, he said, has done a great deal of testing of the weapon in its advanced development stages. In fact, the testing that the Air Force has done is more than is normally performed at this stage of a weapon's development. However, both this testing and that of others working on the same problem have brought to light a number of uncertainties in the operation of the IIR Maverick. These uncertainties lead to the recommendation that tests to resolve them be conducted as quickly as is practical.

The DDT&E went on to say that he intended to work closely with the Air Force in planning tests that could clear up the uncertainties about the weapon's operation. In this regard, he stated the proposal that test plans whose design was specifically intended to resolve operational uncertainties be fully coordinated with and agreed upon by the Office of the Secretary of Defense before the commencement of the testing itself. Nevertheless, the DDT&E concluded, he felt that the IIR Maverick missile system's testing and evaluation had come to such a point that the weapon could be considered ready to enter its next phase of development.

ITEM 4: Operational uncertainties listed by the DDT&E

The following were the uncertainties in the IIR Maverick's operation that the DDT&E acknowledged in September 1976:

--defining the thermal characteristics of the battlefield,
--determining the effectiveness of the aircrews in both single-seat and two-seat aircraft operating in a battle area with simulated thermal characteristics,
--assigning a degree of confidence to laser designations by air and ground forward-area commands on a simulated battlefield,
--determining what the recognition and detection ranges are on a simulated battlefield,
--determining the temperature characteristics of Soviet battle vehicles,
--determining the effects of countermeasures during battlefield operations, and
--evaluating the missile's utility in combat with specific defenses, tactics, and attack profiles.

ITEM 5: Deficiencies of planned Air Force tests

The Assistant Director for Tactical Systems Test and Evaluation stated on October 14, 1976, that the planned Air Force tests of the IIR Maverick would not resolve certain operational uncertainties for the following reasons:

--the test results would be qualitative and variously interpretible because the tests were not set up to measure the timing and positions of aircraft and targets,
--the scenarios did not include the thermal clutter that joint operational tests and evaluations should have,
--the tests would not measure how the weapon operates in thermal clutter with the help of cueing aids,
--lack of instrumentation meant that the tests would not reveal accurate detection ranges in thermal clutter,
--how infrared detection ranges decrease in the presence of aerosol in weather like that of the Federal Republic of Germany would not be measured,
--countermeasures would not be used, and
--the test results would be open to challenge because the tests were not to be conducted by independent agencies.

ITEM 6: The phases of employment of the IIR Maverick

The overall objective of the test was to provide data that would augment the understanding of the operational uncertainties of the IIR Maverick that had been identified by DSARC II's review of the missile program. The missile was to be used as a stand-off, air-to-surface weapon against enemy armor and air defense units behind the forward edge of the battle area in weather and battlefield conditions simulating those of combat in a midintensity conflict in central Europe in 1982. Two scenarios were planned: close air support and preplanned interdiction strikes.

There are two phases in employing the IIR Maverick--navigation and attack. The navigation phase begins when an aircraft leaves an airbase to fly to an initial point. Next, the attack aircraft departs the initial point at the briefed heading, "pops up" at the briefed time and distance, and rolls out at the briefed heading. The "pop up" point is the point of transition from navigation to attack.

The attack phase begins after the rollout, which places the aircraft on a bearing toward the target area. The pilot finds and acquires the target area and begins the "wings-level" part of the pass, during which a target is detected and "acquired." The pilot acquires a target area, transfers the target area to the video display, detects and acquires a target on the video display, "locks onto" the target, and finally launches the IIR missile. ("Acquiring" a target area is defined as using an acquisition or cueing aid to adjust the attitude of the aircraft so that the gunsight reticle--the grid in the eyepiece, or the "pipper"--is on the probable target area.) The "launch and leave" capability of the missile is intended to enable the pilot to turn away after launching the missile and leave the area with evasive maneuvers; meanwhile, the missile is to stay locked on to, and eventually kill, the target that the pilot selected.

In this test, the start of a pass over the battlefield was defined as the time when the pilot reached the initial point, and the end of the pass was the time of the simulated launching. The process of reaching the initial point and the process of "killing" (or not killing) the target were not part of the test.

ITEM 7: Operational uncertainties to be understood

The following is a statement of operational uncertainties in an October 1976 memorandum of understanding on IIR Maverick testing and evaluation between the Assistant Director for Tactical Systems (Test and Evaluation) and the Assistant Director for Air Warfare, both in the Office of the Director of Defense Research and Engineering, and the Director of the Tactical Air Force Division in the Office of the Director of Planning and Evaluation:

"a. What is the thermal character of the battlefield? (What is the thermal image and decay of typical

battlefield events such as munition explosions, killed targets, IR decoys, and dissimilar IR ground texture? How does this IR clutter affect the operator's capability to interpret a true target?)

b. What is the thermal character of proposed targets?

c. For both multi-crew and single-seat aircraft under various day/night conditions:
   (1) What cueing is required?
   (2) What are the typical target detection ranges? This is defined as the range at which targets can be distinguished from background clutter.
   (3) What are target acquisition ranges? This is defined as the range at which the operator can lock onto the detected target.
   (4) What are target recognition ranges? This is defined as the range at which the aircrew, with a high degree of confidence, can distinguish targets; that is, a tank can be distinguished from a truck or hot bomb crater. How is this range degraded in a typical FRG [Federal Republic of Germany] atmosphere?
   (5) Does operator workload degrade IIR Maverick effectiveness when comparing single-seat with multi-crew aircraft?

d. Can a pilot locate tanks at night, given map coordinates of a reported location?

e. How does the use of countermeasures degrade the utility of the system?

f. Can desirable/survivable tactics, or delivery modes, be employed from the point of target detection to weapon release?"

ITEM 8: The JTF's original data analysis plan

The original plan was presented as a list. For brevity, we have run the list on as continuous text, adding punctuation, expanding abbreviations, and supplying other grammatical links as appropriate. Thus, item 8 is a paraphrase rather than a direct quotation and should be so treated in citing to it (see II.C.13).

The test is to assess the operational capabilities associated with the transition from the navigation to the attack phase with the IIR Maverick by day, at night, and with limited visibility. Mission results from each experiment will be categorized, summarized, and compared. The measures of effectiveness will be (1) the probability of acquiring the target area, given departure from an initial point, and (2) the probability of launching on a valid target, given departure from an initial point. For evaluation,

the test results will be grouped, reported, and discussed in order to show the influence of the time of day and weather conditions on mission effectiveness. The relative effect of the time of day on mission success will be determined by comparing results within specific groups of missions. The relative effect of visibility will be determined by comparing results within other specific groups.

The test is to assess the ability of the system to attack typical IIR Maverick targets in an environment representative of a central European battlefield by day, at night, and with limited visibility. Mission results from each experiment will be categorized, summarized, and compared. The measures of effectiveness for all passes, given target-area acquisition, will be (1) target-area acquisition time and range, (2) target detection time and range, (3) target lock-on time and range, (4) missile launch time and range, and (5) the probability of launching on a valid target, given departure from an initial point. For evaluation, the test results will be reported in a way that shows the influence of acquisition aids and target scenarios on mission effectiveness. Because of the small number of discrete experiments and the necessity of varying more than one factor between experiments, it will not be possible to make direct comparisons of variables. The relative effect of having Pave Penny and not having Pave Penny will be determined by comparing results within specific groups of missions.

The test is to assess the ability to employ the IIR Maverick system in single- and dual-place aircraft operations. Mission results from each experiment will be categorized, summarized, and compared. The measures of effectiveness will be (1) the probability of acquiring the target area, given departure from an initial point; (2) for all passes, given target-area acquisition, (a) target-area acquisition time and range, (b) target detection time and range, (c) target lock-on time and range, (d) missile launch time and range, and (e) the probability of launching on a valid target; and (3) the probability of launching on a valid target, given departure from an initial point. The relative effect of single-place A-7 and dual-place F-4 aircraft on mission effectiveness will be determined by comparing results within specific blocks of experiments. If the test reveals functions critical to the successful employment of the missile or operational limitations on its use in either type of aircraft, they will be identified and reported even though a direct comparison is not possible.

The test results are to be reviewed as a function of the meteorological and thermal measurements; the test is to assess the overall operational capability of the IIR Maverick system when it is employed in different weather and with different targets, countermeasures, and battlefield thermal clutter. Because of the test location and resource constraints, the data on operational capability to be acquired during this JOT&E will apply directly to a limited cross-section of weather conditions, target types, and

thermal clutter backgrounds. Extensive data are being collecte on weather and the thermal signatures of target and backgrounds. It is expected that these data elements can be applied to the general characteristics of the IIR Maverick system and to performance models derived from previous testing to project the system's capability in other weather and thermal environments. The measures of effectiveness will be (1) the probability of acquiring the target area, given departure from an initial point; (2) for all passes, given target-area acquisition, (a) target-area acquisition time and range, (c) target lock-on time and range, (d) missile launch time and range, and (e) probability of launching on a valid target; and (3) the probability of launching on a valid target, given departure from an initial point. In the data evaluation, previous reports on the IIR Maverick system's performance will be reviewed, and particular emphasis will be given to reported models of performance as a function of meteorological and thermal environments. The results obtained during the JOT&E will not directly correspond to previous test results because those results were obtained under carefully controlled development test conditions whereas JOT&E testing will be conducted in a more realistic operational environment. Previous test results should give some insight into the effects of meteorological and thermal factors on overall capabilities. Wherever possible, the JTF will report projections of IIR Maverick capabilities in a midintensity European conflict.

The experimental approach for the thermal measurement of foreign armored vehicles will be as follows. Data for the test objective will be obtained primarily from calibrated thermal measurements of Soviet tanks, armored personnel carriers, and air defense units in a variety of environments. This measuring will be conducted by the Night Vision Laboratory independently of the two-sided field test. An important feature of these measurements will be that the armored vehicles will carry baggage, gear, and equipment typical of operational employment. The major variables are solar insulation (hourly measurements), precipitation, target vehicle type, activity state of the vehicle, depression angles, and terrain. As for the instrumentation, some of the measurements may be made at the sites of the two-sided field test in order to obtain measurements comparable to those of the test's actual target vehicles. Each measurement will be made by a static imaging camera system and will be reduced at the Night Vision Laboratory in order to obtain apparent radiation temperatures and differentials with the background appropriate for the IIR Maverick. The plan for acquiring these measurements is given in more detail in a separate appendix to the data analysis plan.

The experimental approach for the thermal measurement of the battlefield will be as follows. Data for the test objective will be obtained from the infrared images of thermal events on the battlefield and of armor operations in various environments, including Europe. This measuring will be conducted independently of the two-sided field test. As for the instrumentation, its details will be specified in addenda to the test plan. However, it is anticipated that this activity will obtain infrared images of the

explosion of general-purpose bombs and artillery rounds so that their appearance can be observed through time in infrared imaging devices. These measurements might be made at various sites where it is possible to achieve such effects. The seeker to be used can be an IIR Maverick seeker or some other infrared imaging equipment, such as the Navy's A-6/S-3 FLIR system or RF-4C/AAD-5 system. No instrumentation other than video tape recording of the images is envisioned, although documentation of the targets or the effects that are observed will be needed. To minimize costs, a helicopter should be considered as the platform for the seeker. However, alternatives such as Navy fixed-wing aircraft carrying thermal imaging devices could be considered if costs are comparable.

ITEM 9: The characteristics of proposed and actual test missions with "good" visibility and "poor" (less than 5 statute miles)

	Visibility		Total	
Time of day	Poor	Good	Number	Percent
Day/midday				
Proposed	3	3	6	33
Actual	2	8	10	44
Night/dusk				
Proposed	3	3	6	33
Actual	2	5	7	30
Night/midnight				
Proposed	0	3	3	17
Actual	1	3	4	17
Night/Predawn				
Proposed	0	3	3	17
Actual	0	2	2	9
Total proposed	6(33%)	12(67%)	18	--
Total actual	5(22%)	18(78%)	23	--

ITEM 10: A comparison of A-10 target-area acquisition ranges (differences significant at .01 on Mann Whitney U Test)

Mission 7
Time = 12:00 noon
Absolute humidity = 5.2
Visibility good: 7 miles

Pass	Range (1,000 ft)
1	
2	
3	
4	
5	
6	
$\bar{X}$	

Mission 20
Time = 2:00 p.m.
Absolute humidity = 5.5
Visibility poor: 1.5 miles

Pass	Range (1,000 ft)
1	
2	
3	
4	
5	
6	
$\bar{X}$	

ITEM 11: Passes with target detection ranges of 500-1,000 feet less than wings-level range a/

Mission type	All passes b/	Passes with target detection range 500-1,000 ft less than wings-level range	
		Number	Percent
A-7 CAS	20		
A-7 PPI	15		
A-10 CAS	36		
A-10 PP/CAS	30		
Total	101		

a/CAS = close air support; PPI = preplanned interdictions; PP = Pave Penny.
b/Data for some passes were not available.

ITEM 12: Number of passes by type of mission a/

Mission type	Target		Abort	No data b/	Countermeasures test	Total
	Valid	Invalid				
A-7 CAS						
A-7 PPI						
A-10 CAS						
A-10 PP/CAS						
Total						

a/CAS = close air support; PPI = preplanned interdiction; PP = Pave Penny.
b/These passes were not counted as "for-the-record" passes.
c/Seven passes from the first mission were omitted because the pilots lacked training.

ITEM 13: IIR Maverick test director's comments on countermeasures report

The memorandum reprinted in its entirety here was addressed to the Office of the Test Director, of the Joint Services Electro-Optical Guided Weapons Countermeasures Test Program, to the attention of DRXDE/TD, at White Sands Missile Range, New Mexico. The subject of the memorandum, written by Major General Howard W. Leaf of the U.S. Air Force, was given as "Final Draft Report of Countermeasures Evaluation of the IIR Maverick JOT&E." There were two attachments, which we have not included, one containing recommended changes and the other containing the distribution list for the memo.

"1. The OTD [Office of the Test Director] Countermeasures Evaluation of the IIR Maverick JOT&E has been reviewed and recommended changes and distribution list are attached for your consideration.

"2. As you are aware, the test results contained in the OTD report differ from those in reports of the Joint Test Force (JTF) and Systems Planning Corporation (SPC), OSD's independent evaluation agency for the IIR Maverick JOT&E. The difference of consequence occurs with respect to the number of record test passes that resulted in simulated attacks on valid targets (tank or APC [armored personnel carrier]) as opposed to invalid targets (all other thermal signatures on the battlefield). Whereas the JTF and SPC assessed valid targets on 59 of the 105 main test passes, the OTD report reflects 55. Although the difference is small, it concerns an important measure of IIR system performance and alters a basic test result. Since as Joint Test Director, I previously signed and published the JTF main test report, it would be inappropriate to co-sign the OTD report containing different test results. I do consider it within your prerogative to publish your report as an OTD document. The precedent to do so has been established by Air Force Studies and Analysis' publication of Annex C, Survivability Analysis, as a stand-alone Air Force document. I recommend this approach for the OTD report.

"3. As you finalize the OTD report, I ask that you make a last analysis of the four test passes in question in a final attempt to harmonize the test reports. In the event your assessments still differ from those of the JTF and SPC, please attach a copy of this letter to your report at distribution."

ITEM 14: Percentages of valid target acquisition in four reports a/

Mission type	JTF	System Planning Corp.	Countermeasures	Air Force
A-7 CAS				
A-7 PPI				
A-10 CAS				
A-10 PP/CAS				
Total				

a/These reports are items 21 (JTF), 23 (System Planning Corp.), 19 (Countermeasures), and 10 (Air Force) in our appendix II, section C. CAS = close air support; PPI = preplanned interdiction; PP = Pave Penny.

APPENDIX IV

ITEM 15: Mean launch and abort ranges by IIR Maverick mission (1,000 feet) a/

Mission type and number	Mean launch range/n		Mean abort range/n	No. of passes not for the record
	Valid targets	Invalid targets		

a/CAS = close air support; PPI = preplanned interdiction; PP = Pave Penny.

ITEM 16: Number of valid and invalid targets with and without Pave Penny in A-10 aircraft in close air support

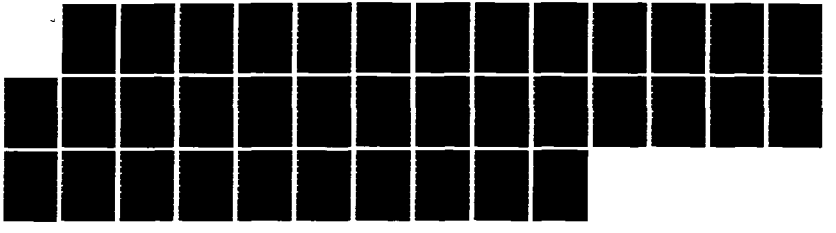
	With	Without	Total
Targets			
Valid			
Invalid			
Aborts			
Total			

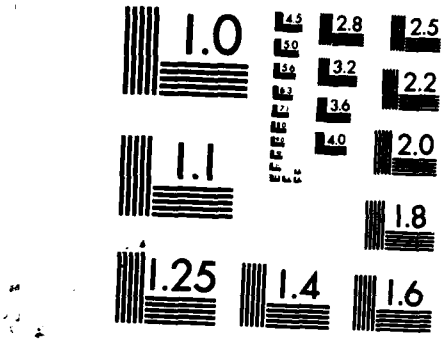
ITEM 17: Simulated launch ranges (1,000 feet)

ITEM 18: Tracking time by pass outcome

AD-A139 427 HOW WELL DO THE MILITARY SERVICES PERFORM JOINTLY IN COMBAT? DOD'S JOINT.. (U) GENERAL ACCOUNTING OFFICE WASHINGTON DC PROGRAM EVALUATION AN.. 22 FEB 84 3/3

UNCLASSIFIED GAO/PEMD-84-3 F/G 15/7 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

APPENDIX IV

ITEM 19: Countermeasure results

ITEM 20:

ITEM 21: Gaps in the thermal data

The System Planning Corporation reported that

"Among the more important gaps in the data needed to predict thermal conditions are:

--Thermal parameters characterizing how the radiant temperature of a T-62 varies with solar insolation [sic] are not known.

--Thermal parameters characterizing how the radiant temperatures of European backgrounds vary with solar insolation [sic] are not known outside of the single background at Grafenwoehr.

--The equilibrium radiant temperatures of European backgrounds and the T-62 tank for various conditions are not known. These temperatures, along with the thermal parameters, are needed for reliably producing signatures of threat targets in Europe.

--No historic data are known to exist documenting concurrent hourly solar insulation [sic] and atmospheric transmittance sampled over several seasons at a European location."

TECHNICAL DATA FOR CHAPTER 5

ON TASVAL

The items in this appendix are supplementary and supporting technical data for the presentation of our analysis in chapter 5. The items are keyed by number to the text, where they are cited as parenthetical references to this appendix. They are also listed immediately below, each with its item number and a short descriptive title of its contents.

ITEM		Page
1	Chronology of the TASVAL JT&E	178
2	Description of TASVAL's objectives and design	179
3	Near-real time casualty assessment	182
4	Force composition	183
5	Training	184
6	Normalization and validation procedures	186
7	Mean number and percent of friendly aircraft engagements against enemy forces per trial	189
8	Mean number of expected enemy casualties per trial from air-to-ground engagements with the Maverick missile and the GAU-8 gun	189
9	The classification of events for defining probability of kill	190
10	Friendly air force firings and pairing rates	190
11	Fixed assumptions and variable factors in air-to-ground probability-of-kill models	191
12	Mean number of fully valid enemy weapon systems' ground-to-air engagements and their expected contribution to friendly air force attrition per trial	191
13	Mean expected aircraft attrition per trial	192
14	Resolved and unresolved ground-to-air and air-to-ground engagements by team	192
15	Enemy air defense unit firings and pairing rates	193
16	Mean number of air-to-ground engagements against dead targets per trial by team	193
17	DOD comments on TASVAL's usefulness	193

ITEM 1: Chronology of the TASVAL JT&E

May 1977	IDA is asked to prepare preliminary test design
June 1977	DDT&E is briefed on test design
August 1977	Services are briefed on test design
September 1977	Memo from Under Secretary of Defense for Research and Engineering requests JT&E to be conducted April-June 1978, with a preliminary report on July 1978 and the final report on September 30, 1978
October 1977	Commander of the U.S. Army Operational Test and Evaluation Agency is appointed as joint test director

December 1977	IDA publishes test design; draft JTF test plan is presented at conference at U.S. Army Operational Test and Evaluation Agency; decision is made that credible test could not be conducted before April 1978
January 1978	Second draft JTF test plan is presented at conference at U.S. Army Operational Test and Evaluation Agency
January 30, 1978	Second draft test plan is circulated to services; Ft. Hood is selected as test site; test is rescheduled to begin in April 1978
February 20, 1978	Services comment on test plan returned
March 1, 1978	Test is delayed until spring 1979 because of instrumentation problems, lack of adequate threat simulators, and need to change test site
May 1978	Ft. Hunter Liggett is designated as new test site
October 1978	Joint test director is appointed
November 1978	Joint task force is formally established
Mid-January 1979	All members of JTF are assigned
March 1979	Test is scheduled to begin but delayed because of wet ground and delays in developing multiple computer system
	Major units and JTF staff are moved to Ft. Hunter Liggett
	Final test plan is published
May-June 1979	Joint test director decides to use SEL-86 computer instead of MCS
May 16-August 2, 1979	Exploratory trials are conducted
August 3, 1979	Technical advisory board gives OK
August 8-September 27, 1979	Record trials are conducted
May 1980	JTF publishes test report
June 1980	U.S. Air Force Studies and Analyses publishes test report
December 1980	IDA publishes test report
April 1981	U.S. Army publishes test report

ITEM 2: Description of TASVAL's objectives and design

The purpose of TASVAL was to find out how many aircraft would be lost and what targets could be killed in typical close air support missions in a heavily defended area of central Europe. The original test objectives were as follows:

--determine the rate of loss among friendly aircraft attacking enemy antiarmor defenses in moderately to heavily defended areas and determine which enemy weapons or combinations of weapons most effectively destroy friendly aircraft;

--determine the rates of damage and destruction among enemy armored targets and determine which attack aircraft,

weapons, and tactics result in the highest rates of target destruction;

--determine the relation between the loss of friendly aircraft and the destruction of armored targets, given different tactics and combinations of friendly attack forces;

--evaluate the "synergistic" effect of the combined operation of AH-1S and A-10 aircraft on "kill" and "survivability" rates;

--evaluate the effects of weather, given assumed ceilings and visibility restrictions, and electronic warfare countermeasures on the ratios of friendly aircraft losses to enemy armored target losses.

On August 23, 1979, some 15 days after the record test trials started, the DDT&E revised these test objectives (to the three listed in figure 32 in chapter 5).

The first significant change was the deletion of the fifth objective--evaluating the effect of weather and electronic warfare countermeasures--because there was no satisfactory way to simulate European weather and because considerations of security and the environment prevented all active electronic warfare countermeasures except communications jamming. The second significant change was from a quantitative to the qualitative evaluation of the four remaining objectives. This change was made because it was believed that simulation models are inherently limited in providing meaningful absolute "probabilities of kill" and that the discrete data provided by the instrumentation systems made for uncertainties in the flight paths.

According to the test concept, TASVAL would simulate ground-to-air defense activity by Warsaw Pact forces against NATO forces at a military front near Fulda, a city northeast of Frankfurt in West Germany. The JTF specified only one independent variable--that is, "strike package," or combination of aircraft being assessed--but the test site valleys (Gabilan and Nacimiento) were also used in the analysis and the reports of test results as an independent variable. The test concept further postulated that the conflict to be simulated would be a conventional, nonnuclear war using weapon systems available by December 31, 1980.

TASVAL was conducted at Ft. Hunter Liggett, California. The site at the Gabilan Valley was about 6 kilometers wide with ridges 400-500 feet high on the northeast side and ridges 700-1,000 feet high on the southwest. The Nacimiento Valley site consisted of two valleys side by side, Nacimiento and Stony, with a combined width of approximately 5 kilometers but separated by a ridge 400-500 feet high. The floor of Nacimiento in kilometers measures about 2 by 8, Stony Valley about 1 by 6. A 600-700-foot-high ridge borders Nacimiento and Stony on the northeast.

Gabilan and Nacimiento each provided approximately 8 square miles for ground maneuvers. This area was instrumented with a range measuring system composed of three types of field equipment that was able to specify the position of all players, transmit information about events to a central computer, and transmit commands and information back from the central computer to the players. The three types of equipment were (1) communication towers, called "A-stations," for measuring the distances between players and providing a data link between (2) transponders, called "B-units," mounted on each player and providing the basis for range measurements, and (3) a "C-station," which controlled the range communications of the players' B-units and sent to the central computer data on locations and events. The central computer, the SEL-86, recorded the data, assessed ground casualties and "near-real" time, and sent these data on to the test operations center. This center displayed graphically the locations of all active players, all firings, and all hits for the test control personnel.

TASVAL was one of the most complicated two-sided field tests ever attempted by the JT&E directorate. It involved more than 100 players, each with instruments for documenting positions, firing, hits, and firer-to-target pairings. "Near-real" time casualty assessment was provided for each trial in ground-to-ground and air-to-ground but not ground-to-air engagements (see item 3 in this appendix for a description of near-real time). Tanks and other ground weapon systems that were destroyed were so marked by the data collectors, who threw a purple smoke grenade beside them and required the crew to stop and to cease engaging targets. The training was extensive (see item 4 in this appendix for a description of it).

Thirty-five exploratory trials were conducted between May 16 and August 2, 1979, and actual test trials were conducted from August 8 through September 27, 1979, some 15 months after the date originally proposed for completion. The test trials generally included the friendly "Blue" aircraft providing support to friendly Blue ground forces, which were defending against enemy "Red" ground forces conducting a breakthrough attack. A second scenario, originally planned to represent the Blue attacking the Red defenders, was dropped from the test. A typical test trial was made up of the following activity. First, there was a reconnaissance period. Generally, the enemy ground forces contained about 80 vehicles, which included 24 air defense units. As the enemy proceeded toward the defensive position of the friendly forces, made up of about 14 units, the friendly air and ground forces tried to prevent the enemy forces from meeting their objective and to drive them into a hasty defense. As for tactics, the friendly forces could lethally suppress enemy air defense units with artillery and other ground-to-ground fire and with TOW, Maverick, and GAU-8 air-to-ground fire. Trials typically lasted minutes, ending when friendly strike aircraft had used up all available weapons or left the area.

The three combinations of aircraft, or "strike packages," were as follows. First, close air support was provided by attack helicopters in conjunction with scout helicopters. Second, close air support was provided by Air Force A-10's (typically starting farther away from the battle area than the attack helicopters) in conjunction with scout helicopters and O-2 aircraft. Third, close air support was provided by both attack helicopters and A-10's as a joint air attack team, the Army and the Air Force operating together in the same air space.

Of the 58 record trials that were conducted, 45 were declared fully valid by the validation procedures described below, in item 6 in this appendix. In the analysis, the three revised objectives on effectiveness, attrition, and synergism were stated so as to require computer simulations of the test data. For the effectiveness objective, for example, instead of using test-related results such as the number of enemy vehicles detected or the amount of simulated fire attempted per trial or the time and standoff ranges of each detection and simulated firing event, the analysis used the number of enemy vehicles destroyed per trial. As they were stated, the objectives led to conclusions being drawn from models of the data rather than from the data themselves.

ITEM 3: Near-real time casualty assessment

"Real" time is defined as the time in which a firer engages a target during battle, including the making of all removal decisions and the providing of all associated cues. It stops before any player becomes involved in any other engagement. "Near-real" time refers to the fact that the data about an engagement--who the firer is, what types of weapons are used, what the target is, how far away the firer is from the target, and so on--are gathered in real time, not longer than one second after the engagement, and are extracted and assessed some few seconds after the impact and detonation of a projectile, depending on its type, because of the time it takes to compute and transmit the data. The limitations of the computer meant that the performance of the aircraft could not be measured accurately and fed into the models in real time.

For TASVAL, there were four levels of assessment of casualties in near-real time: the assessment of casualties from data (1) collected not in real time but in post-trial data reduction and processing and from data (2) extracted from near-real time engagements involving ground-to-ground activity, (3) ground-to-ground and air-to-ground activity, and (4) ground-to-ground, air-to-ground, and ground-to-air activity. The specific objectives of assessing casualties in near-real time were to

"(1) Generate data on the effects of real time attrition on the effectiveness and survivability of attack aircraft and air defense weapons.
(2) Induce players to employ prudent tactics.
(3) Motivate gunners to engage targets aggressively.

(4) Allow real time tactics to be responsive to the dynamic battle situation.
(5) Shape the battle so that ground players are prevented from making unrealistic geographical advances across the battle area.
(6) Preclude aircrews from unrealistically exposing their aircraft for extended periods and preclude unrealistic attack profiles." (II.D.13, pp. 1-9 and 1-10)

ITEM 4: Force composition

Friendly air					
Attack helicopter		A-10		Joint air attack	
Aircraft	Ordnance	Aircraft	Ordnance	Aircraft	Ordnance
5 AH-1S	8 TOW	4 A-10	6 EO Maverick, 1,350 rounds GAU-8	4 A-10	6 EO Maverick, 1,350 rounds GAU-8
3 OH-58	None	1 OH-58	None	5 AH-1S	8 TOW
		1 O-2	None	4 OH-58	None
				1 O-2	None

Enemy ground		Friendly ground
Threat	Simulator	Equipment
31 T-72 tank	31 M-60 tank	10 M-60 tank
10 BMP SAGGER	10 M-220 APC, TOW	2 M-220 TOW
6 122-mm SP howitzer	6 M-60 tank	2 APC M-113 APC
12 BMP, SA-7	12 jeep, trailer, SA-7	
4 SA-8 fire unit	2 Hawk battery, each with 2 HIPIR, 2 TGT	
4 SA-9 fire unit	4 improved Chaparral	
4 ZSU-23-4 fire unit	3 ZSU-23 vehicle, 1 ZSU radar van, 1 nonfire TGT vehicle	
6 BTR-60 command vehicle	6 M-113 APC	
6 ammunition vehicle	6 M-35 truck	

ITEM 5: Training

In this item, we describe the training for TASVAL that was completed before the exploratory trials--that is, between March 12 and April 20, 1979--and during the exploratory trials--from May 16 to August 2, 1979. For brevity, we have paraphrased the JTF's appendix G outline (II.D.7), running it on here as continuous text.

The main objectives of the training program were to orient all TASVAL participants to Ft. Hunter Liggett and train them in safety and security procedures and to give the players and data collectors a working knowledge of the vehicular instrumentation and cueing devices. The orientation, which the JTF conducted, included a welcoming address by the joint test director and the post commander, an orientation to Ft. Hunter Liggett, a statement of the test's concepts and a schedule, a presentation of the test scenario and sequence of events, some explanatory material on data collection and reduction, and information on test safety and operations security.

The ground players were trained in instrumentation in a 2-hour overview of the entire system and the interaction of its elements and, for vehicle crew members, specialized instrumentation training on their specific vehicle, weapon, and associated hardware. A 2-hour, hands-on training session was given in devices to be used as cues to enemy equipment to the following groups:

Ground force	Simulated enemy	Cueing device
T-72 crews	M-60A1 tank	Hoffman
SP122 crews	M-60A1 tank	Hoffman
BMP/SAGGER crews	M-220 TOW carrier	Frankford Arsenal
SA-7 gunners	XM-76	Distress flare
SA-9 crews	Chaparral	ATWESS
SA-8 TEL drivers	2-1/2-ton truck (modified)	ATWESS
M-60A3 crews	M-60A1 tank	Hoffman
M-220 crews	N/A	Frankford Arsenal

A total of 49 trials were conducted before record trials began, in order to give the ground players additional training.

The attack helicopter teams were trained in instrumentation in a 2-hour overview, followed by specific instruction on the AH-1S and OH-58, and in 1.5 hours of instruction on the operation of the ATWESS cueing device. The 7-17 Air Cavalry were briefed on the local flight area in 6 hours of instruction that included presentations by the Ft. Hunter Liggett aviation officer, the Ft. Ord aviation section, the 155th Attack Helicopter Company, and the TASVAL air operations section. The teams were given opportunity to fly over the area in order to become familiar with the terrain and possible safety hazards, and 8 formal exploratory trials were conducted to give them opportunity to develop and refine their tactics and procedures. Twelve additional trials that were to become record trials were conducted, but they were not finally

counted as record trials because of instrumentation and operational problems, although they were useful as training.

The A-10 aircrews were trained at Nellis Air Force base, Nevada, by the Tactical Fighter Weapons Center. In the vicinity of Tolicha Peak, near the base, they flew against threat simulators in order to refine their tactics and techniques for using the Maverick and the GAU-8 weapon systems. Because of a shortage of aircraft and the inclement weather, only 47 of the 61 sorties that were flown were considered effective. While 12 four-ship missions were flown, only 5 were considered effective. When aircrews were not scheduled to fly, they monitored the day's training missions on the ground in the threat simulators. Following this tactical training, the A-10 aircrews were deployed to Ft. Hunter Liggett so that they could become familiar with the area, and all future tactical training was conducted at that complex. They were also briefed at Ft. Hunter Liggett on joint air attack tactics for range orientation through the use of the Cobra and Scout helicopters. Seven formal exploratory trials were conducted with the A-10 strike package, which gave the crews further familiarization with the area and helped them refine their tactics, coordination, and procedures, and 32 instrumentation missions were flown with the same, if much more limited, purpose.

Training in joint air attack coordination consisted of meetings between the attack helicopter crews, the A-10 pilots, and their commanders to discuss joint flight operational procedures, communications, and general interaction followed by flights over the Ft. Hunter Liggett local flight area so they could become familiar with the terrain, communications, procedures, and tactics. There were 20 formal exploratory trials, which gave additional opportunity for the development and refinement of tactics, coordination, and procedures.

Airborne forward attack coordinators were oriented at Nellis Air Force Base in TASVAL safety, tactics, and the local area and were given flight training in the vicinity of Tolicha Peak. To develop proficiency in controlling the pop-up attacks of the A-10's, they practiced computing initial-point-to-pop-up-point parameters. Upon deployment to Naval Air Station Lemoore, they executed several familiarization flights over the Ft. Hunter Liggett complex, where they were also given TASVAL briefings on the ground. Their services were required in 27 formal exploratory trials with the A-10's, which increased their familiarity with the area and helped them refine their coordination and procedures.

Data collectors received the same training in instrumentation and cueing devices as the ground players, being integrated with their classes whenever possible. The data collectors were also given 10 hours of instruction on data collection procedures, and refresher and remedial classes were provided as required.

The "Red" or "enemy" ground and air defense units were given tactical threat training by the Army Forces Command Training

Detachment. Their training included lectures, practical and table exercises, tactical exercises without troops, and company and battalion field exercises.

ITEM 6: Normalization and validation procedures

IDA's data normalization procedures were as follows:

"the number of engagements that were unassessed is not inconsequential. Furthermore, the assessment rate varies by system and by strike package-valley combination. Comparing weapon systems only on the basis of assessed engagements is invalid because of these variations in the completeness of the data base. This is especially true in computing exchange ratios, since air-to-ground engagements were assessed more frequently than ground-to-air engagements.

"A normalization procedure was developed to adjust for variations in the completeness of the data base. The effectiveness of a particular weapon when used against a certain target type* was established using the assessed firings. This was done for each strike package-valley combination.

"Using the average Pk [probability of kill] per assessed firing, three measures of performance were calculated. First, estimated losses due to paired and assessed firings were calculated by summing the Pk of those engagements that were assessed. Secondly, estimated losses for all paired launches were calculated, taking into account those firings that were paired but not assessed. This measure was obtained by multiplying all paired launches by the average Pk per assessed firing. Lastly, estimated losses due to all valid firings were calculated, again using the average Pk per assessed launch. This last measure is the best estimate of the system's performance after adjusting for the lack of completeness of the data base.

"This normalization process assumes implicitly that the unassessed and unpaired firings had the same distribution of Pk values as the assessed firings. There is no way to prove this assumption but it is probably closer to the truth than the assumption that all unassessed firings have a Pk equal to zero.

"*For the ADU firings, each A-10, AH-1S, and OH-58 aircraft was considered separately. For air-to-ground firings, all ground targets were considered one type of target." (II.D. 28, p. 133)

For brevity, we have paraphrased the TASVAL validation committee's procedures, outlined in appendix F of II.D.15. The

TASVAL test data were validated before they were released for analysis. The validators were a committee of five who were directly responsible to the joint test director: one from IDA (the chairman of the committee), one from the JTF analysis division, and one each from the Army, Air Force, and Marine Corps. They reviewed the data on engagements, trials, and post-trial probabilities of kill in three phases.

According to the scenario in the test design plan, a valid trial required a specified number of each type of player with opposing units following characteristic tactics and doctrine. Thus, the first phase of validation addressed whether these validation criteria were met during the field trials. That is, the following criteria were looked for: Was the correct number of players present? Were the tactics representative? Did nontactical factors greatly affect the play of the trial?

A subcommittee was formed to observe each trial from an approprite site and to describe its nature on data forms. Immediately following each trial, the subcommittee met with the field execution division to compare notes and to write a summary of the trial, noting any unusual circumstances in the trial and recommending to the committee whether the trial should be considered valid. Data were provided on the following: environmental factors (smoke, dust, haze), the unusual behavior of any element or unit, the number and type of systems operable at the start of the trial, the number and type of systems that failed mechanically during the trial or were administratively deleted, the number and type of battle casualties, the compliance of the enemy force and the friendly ground force with current doctrine and tactics, the conformance of rotary-wing and fixed-wing aircraft to current doctrine and tactics, air space violations, and the number of enemy and friendly artillery missions submitted and fired. The committee reviewed the subcommittee's report and forwarded its own recommendations, along with any dissenting reports, to the joint test director, who either approved the recommendation or directed that other appropriate action be taken.

The second phase, the data reduction, began when the joint test director had judged a trial valid. Data were then collected from computers, video tapes, photographs, voice recordings, and the field forms if they satisfied the following criteria:

--Could the players' firing events involving aircraft be reproduced from the data at the rate of 40 percent for the SA-7, 60 percent for the SA-8, 60 percent for the SA-9, and 50 percent for the ZSU-23-4?

--Were the data of sufficiently high quality to permit the operation of all ground-to-air "flyout" models?

--Were the results from the ground battles reasonable and consistent?

--Were the ground-to-ground and air-to-ground probabilities of kill reasonable and consistent?

When a "flyout" could be produced from these data, other data were added to it: a list of all engagements supported by completed data on time, the firer, the target, the range (critical illumination period for TOW and SAGGER), the quality of SCORE (Simulated Combat Operations Range Equipment) data and the time-space position file-indicator during air defense engagements, the recommendations on validity, and imperfections in the data (such as why a target could not be identified); the percentages of time each SCORE-pod (a measuring device on the outside of the aircraft) did and did not make contact with ground instruments; the time when raw "position location" data were usable; summary information on the weapon, including the number of times players fired, paired, and were "killed"; summary information on the trial including the percentage of pairings by weapon type and total enemy and friendly ground losses in real time; and several data files.

With these data, the validation committee determined the degree to which the important events in the trial could be accurately reproduced. A trial summary was forwarded to the joint test director, highlighting the subcommittee's report, pointing out abnormalities discovered during data collection and reduction, making a recommendation on validity, and including any dissenting reports. The director then declared either that the trial was valid through the second phase, and that ground-to-air probability-of-kill should be generated for each engagement, or that the trial was invalid and the data should be stored.

The third phase consisted of the final data review. Data that had been declared acceptable through the first two phases were "scrubbed" to resolve uncertainties. A probability-of-kill value was determined for each ground-to-air engagement that had not been previously assessed. A final review was made for reasonableness.

In no case were events or trials invalidated simply because a judgment had been made that the probabilities of kill were unacceptable. The committee determined either that the values were reasonable and consistent or that they were questionable, specifying its reasons for concern and returning the data for investigation and recalculation. Upon the resolution of remaining issues with data management, the validation committee submitted its recommendations on validity, along with any dissenting reports, to the joint test director, who made a final declaration on whether the trial was or was not valid and directed that appropriate action be taken.

ITEM 7: Mean number and percent of friendly aircraft engagements against enemy forces per trial a/

Team and site	Engagements against: Air defense Mean	%	Armored units Mean	%	All enemy vehicles Mean	%
Attack helicopter						
Gabilan	7.6	22	26.9	78	34.4	100
Nacimiento	1.4	4	31.3	96	32.7	100
A-10						
Gabilan	6.0	18	26.2	81	32.2	100
Nacimiento	1.0	5	18.1	95	19.2	100
Joint air attack						
Gabilan	9.8	17	48.9	83	58.5	100
Nacimiento	4.4	8	53.3	92	57.4	100

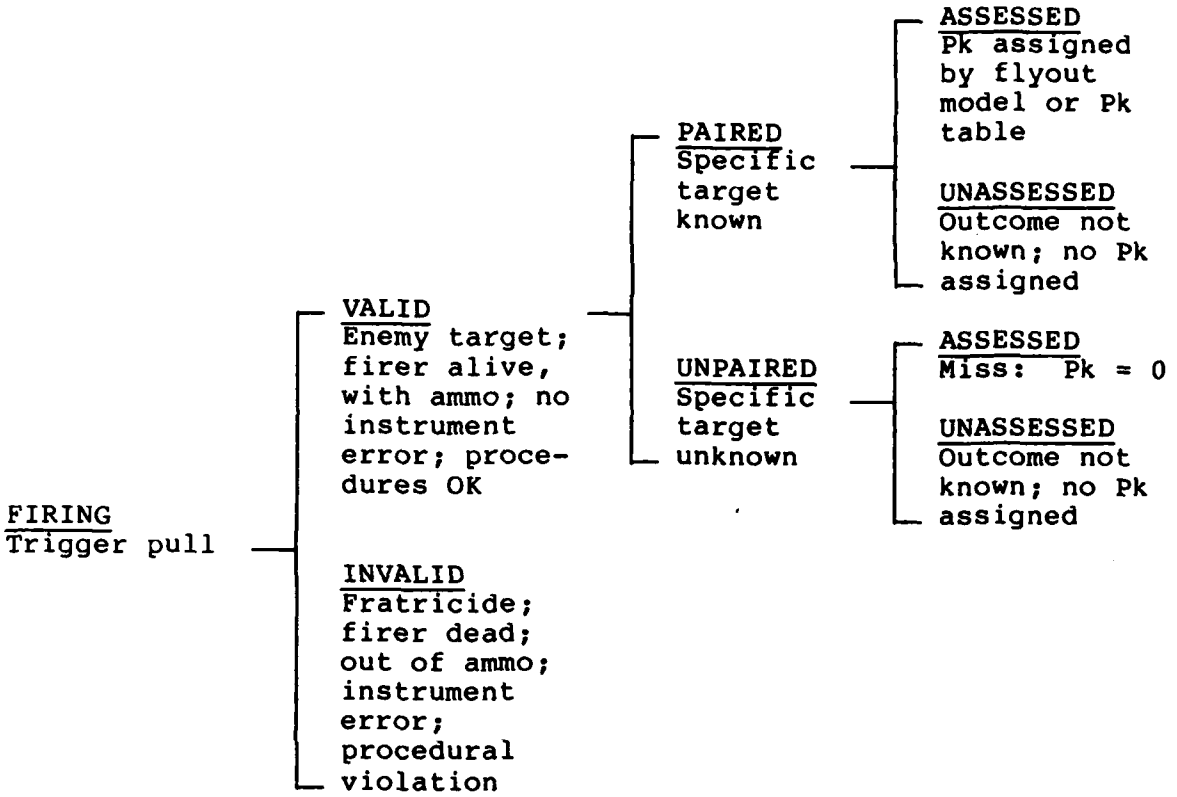
a/"Engagement" is a pairing between a firer and a target. The sum of engagements against enemy air defense and enemy armored units may not add to the total of engagements against all enemy vehicles because of rounding.

ITEM 8: Mean number of expected enemy casualties per trial from air-to-ground engagements with the Maverick missile and the GAU-8 gun

Team and site	Air-to-ground engagements	"Kills" with A-10: Maverick	GAU-8	Total
Attack helicopter				
Gabilan				
Nacimiento				
A-10				
Gabilan				
Nacimiento				
Joint air attack				
Gabilan				
Nacimiento				

APPENDIX V

ITEM 9: The classification of events for defining probability of kill (Pk)



ITEM 10: Friendly air force firings and pairing rates

Aircraft and weapon	Firings in 43 valid trials	Percent paired with targets identified	
		By all means	With manual adjustment
AH-1S and TOW a/			
A-10 and Maverick			
A-10 and GAU-8			

a/AH-1S is the attack helicopter.

ITEM 11: Fixed assumptions and variable factors in air-to-ground probability-of-kill models

Weapon	Target	Fixed assumptions	Variable factors
GAU-8	Enemy vehicles	Target fully exposed, stationary, aspect angle at arithmetic average	Target type; range 1,000-10,000 ft; kill criteria; aspect angle 0-315°; burst length; tracking error
Maverick	Enemy vehicles	Target fully exposed, stationary	Target type; range in terms of fuse; kill criteria
TOW	Enemy vehicles	Target 82 percent exposed, 15 percent stationary, aspect angle at cardioid average for all but T-72	Target type; range 500-3,750 m; kill criteria; aspect angle 0-180° for T-72 only

ITEM 12: Mean number of fully valid enemy weapon systems' ground-to-air engagements (mean) and their expected contribution to friendly air force attrition (percent) per trial

Team, aircraft, and site	SA-7		SA-8		SA-9		ZSU-23-4		Non-ADU a/	
	Mean	%	Mean	%	Mean	%	Mean	%	Mean	%
Attack helicopter										
AH-1S		3		23		24		14		36
Gabilan	3.0		7.7		1.4		4.6		7.4	
Nacimiento	0.3		2.7		0.4		4.7		0.4	
A-10										
A-10		0		76		24		b/		0
Gabilan c/	14.8		15.0		13.5		34.4		0.3	
Nacimiento	10.1		9.3		10.1		24.4		0	
Joint air attack										
AH-1S		10		25		12		5		48
A-10		0		83		17		0		n.a.
Gabilan	12.0		14.6		10.5		16.9		4.5	
Nacimiento	8.1		8.0		6.0		13.8		1.3	

a/Non-ADU = T-72, SAGGER, and 122-mm SP howitzer.
b/Less than 0.5 percent.
c/Non-ADU engagements against OH-58 only.

ITEM 13: Mean expected aircraft attrition per trial a/

Team and site	Aircraft AH-1S	OH-58	A-10
Attack helicopter			
Gabilan	1.6	0.5	n.a.
Nacimiento	0.2	0.5	n.a.
A-10			
Gabilan	n.a.	0.1	2.5
Nacimiento	n.a.	0.2	0.2
Joint air attack			
Gabilan	1.1	0.3	1.6
Nacimiento	0.2	0.4	0.2

a/The AH-1S is the attack helicopter; the OH-58 is the scout helicopter.

ITEM 14: Resolved and unresolved ground-to-air and air-to-ground engagements by team

(U)Resolved engagements were valid trial events for which the outcome could be determined and a probability of kill could be assigned. Unresolved engagements were valid trial events for which no outcome could be determined and no probability of kill could be assigned. The total for ground-to-air includes the approximately 5 percent of all ground-to-air engagements that may have been invalid because the SA-8 target did not meet the prescribed azimuth limits or the aircraft had position location noise before the last 2.5 seconds before missile impact.

Engagement	Attack helicopter Total	Percent	A-10 Total	Percent	Joint air attack Total	Percent
Ground-to-air						
Resolved	229	62	1,001	63	765	49
Unresolved	136	37	519	32	713	46
Total	367	100	1,603	100	1,563	100
Air-to-ground						
Valid and resolved	470	91	392	89	927	94
Unresolved	44	9	49	11	64	6
Total	514	100	441	100	991	100

ITEM 15: Enemy air defense unit firings and pairing rates

Unit	Firings in 43 valid trials	Percent paired with targets identified	
		By all means	With manual adjustment
SA-7			
SA-8			
SA-9			
ZSU-23-4			
Total			
Average			

ITEM 16: Mean number of air-to-ground engagements against dead targets per trial by team

Aircraft and weapon	Attack helicopter	A-10	Joint air attack
AH-1S and TOW a/	6.3	n.a.	10.3
A-10 and Maverick	n.a.	3.6	5.4
A-10 and GAU-8	n.a.	2.4	1.5

a/AH-1S is the attack helicopter.

ITEM 17: DOD comments on TASVAL's usefulness

In this item, we quote three memorandums, all addressed to the Deputy Director for Defense Test and Evaluation. The first is from the Deputy Assistant Secretary for General Purpose Programs; the second is from the Army's Deputy Assistant Secretary for Research and Development; the third is from the Air Force's Acting Assistant Secretary for Research, Development, and Logistics.

Memorandum dated May 1, 1981

"I share your concern regarding our ability and inclination to learn from TASVAL. From the outset we argued against using the test to answer specific time sensitive questions regarding individual systems. As it turned out the test program was compressed to the point that only 45 valid trials were run. While we would have preferred a test such as we had advocated originally--a more thorough examination of the total CAS [close air support] mission area in a more deliberate manner--I would agree that we must build on this start. TASVAL, however, was not without payoff:

--It was the most complicated field test ever attempted and much has been learned, particularly with respect to instrumentation requirements (and limitations), for a test of this scope, and

--Much was gleaned from the test in the context of tactics, operational concepts, and training. These areas are the jurisdiction of the Services and should appropriately be addressed by them.

"While it may be inappropriate to extrapolate raw TASVAL data to make force structure, force mix, or weapons systems acquisition decisions, there are unmistakable insights to be gained from the work. You cite some potential issues that are suggested by the TASVAL results. I believe these are valid and to them I would add:

--Should the Army and Air Force reapportion the CAS mission area, i.e., Army--CAS; Air Force--BI [battlefield interdiction]?

--FAC-X [forward air controller, either]: fixed wing or helicopter?

--How do we improve target acquisition and attack without increasing vulnerability?

"Finally, I recommend that we plan for an expanded test of the CAS mission area on the TASVAL mold.

"Thomas P. Christie
Deputy Assistant Secretary
(General Purpose Programs)"

Memorandum dated May 12, 1981

"The Joint Test Tactical Aircraft Effectiveness and Survivability in Close Air Support Antiarmor Operations (TASVAL) test has provided valuable insights about equipment, tactics, and training. In addition, the TASVAL test advanced the state-of-the-art in many areas of operational testing methodology, instrumentation, and execution.

"The Army has not yet assimilated in detail the major findings of its recently completed independent evaluation of TASVAL; however, an initial review indicates that our present course regarding scout and attack helicopter system development, doctrine, and training is reaffirmed with little modification. A briefing of the major findings of the evaluation is being provided for the Army Staff and will be presented to the CSA in the near future.

"The limitations mentioned in your memorandum did prevent satisfaction of the original TASVAL objectives; however, valuable indications regarding the survivability and effectiveness of the A-10 and the attack helicopter team have been derived. TASVAL also reinforced our doctrine for employment

of scout helicopters in the security and reconnaissance roles and the scout's importance in air defense warning and deception. We confirmed the necessity of the laser range-finder for greater AH-1S effectiveness even when the AH-1S becomes the threat posed by opposing force tanks and other non-air defense weapons was again highlighted as a consideration in attack helicopter team operations; and, it confirmed the need for a longer range, faster, multiple engagement missile.

"For the Army, the scope and realism of TASVAL represents the successful culmination of over ten years of operational testing and experimentation that have influenced the present doctrine, tactics, and ongoing procurement programs for the scout and attack helicopter. At Tab A [not reprinted here] are the Army's tactical observations from TASVAL which have been briefed to the training and doctrine community. The Army is not presently contemplating any specific new procurement actions based solely on TASVAL findings; however, we are making some changes in tactical procedures and training emphasis. Training for scout and attack helicopter pilots is being combined in part for maximum cross training benefit; we are working with the US Air Force to revise the joint manual to improve and standardize doctrine, tactics, and training in Joint Air Attack Team (JAAT) procedures; and we are actively exploring the retrofit of a laser range-finder for all AH-1S attack helicopters.

"There are many meaningful lessons to be learned from TASVAL about conduct of major operational tests. Many 'firsts' were attempted in TASVAL and much was learned which has led to major improvements in the Services' operational testing capabilities. As an example, great advances were made in the development of a Near Real Time Casualty Assessment and Removal (NRTCAR) capability, an essential element in any force-on-force operational testing effort. Army lessons learned regarding testing are contained in Tab B [not reprinted here]. The application of these lessons is presently paying dividends in the Army's testing of the AH-64, Phase II of the Joint Service Test (EW/CAS), and J-CATCH.

"The Army feels that it may be premature to provide a position on conduct of USAF close air support, or changes in Army attack helicopter team tactics or weapons based on the results of TASVAL. As specific answers to your questions are formulated by the Army, our plans for action based on an evaluation of the TASVAL results will be provided to your office.

"Amoretta M. Hoeber
Deputy Assistant Secretary of the Army
(Research and Development)"

APPENDIX V

Memorandum dated May 22, 1981

"James E. Williams, Jr.
Acting Assistant Secretary of the Air Force
(Research, Development, and Logistics)"

TECHNICAL DATA FOR CHAPTER 6

ON ACEVAL

The items in this appendix are supplementary and supporting technical data for the presentation of our analysis in chapter 6. The items are keyed by number to the text, where they are cited as parenthetical references to this appendix. They are also listed immediately below, each with its item number and a short descriptive title of its contents.

ITEM		Page
1	Chronology of the ACEVAL JT&E	197
2	Description of ACEVAL's objectives and design	198
3	Air combat maneuvering instrumentation	202
4	Trial validation	203
5	Overall loss rates and exchange ratios	205
6	The effect of disengagement on kill and loss rates	206
7	Modeling problems	206
8	Missile activity against targets	207
9	F-15 learning as expressed in exchange ratios	207
10	Reported effects of weather on F-14 kill ratios by force ratio	207
11	Other uses of ACEVAL's results	207

ITEM 1: Chronology of the ACEVAL JT&E

May 1974	DDT&E requests the Weapon System Evaluation Group and the Institute for Defense Analyses to perform a feasibility study (also known as "design definition" study)
September 1974	Feasibility study is published
April-May 1975	Deputy Secretary for Research and Engineering establishes ACEVAL and charters the joint test force
December 1975	Institute for Defense Analyses publishes a test plan
August 1976	JTF publishes a test plan
October 1976	AIMVAL/ACEVAL training and pretesting begin at Nellis Air Force Base, Nevada
November 1976	Air Force and Navy approve the test plan
December 1976	Pretest is completed; major changes are made to test plan
January 1977	AIMVAL test trials begin
March 1977	F-14 ACEVAL training trials are conducted
May 1977	F-15 ACEVAL training trials are conducted; changes are made to ACEVAL test plan based on lessons learned in AIMVAL
June 1977	AIMVAL test plan trials are completed; ACEVAL test trials begin
September 1977	JTF publishes its final report on AIMVAL

November 1977	ACEVAL test trials are completed
February 1978	JTF publishes its final report on ACEVAL
January 1979	Institute for Defense Analyses publishes its independent assessment of the ACEVAL results

ITEM 2: Description of ACEVAL's objectives and design

The overall goal of ACEVAL was to answer a question that is central to all military planning: In exactly what way does the outcome of a fight depend on how many combatants are fighting? Thus, the controlled, or independent, variables in the test's design included the type of aircraft, the number of aircraft (both "encounter size," or the actual number of aircraft on each side, and "force ratio," or the ratio of friendly aircraft to enemy aircraft), and the availability of ground control intercept (GCI) information (that is, information telling a pilot the relationship between the aircraft's position and the source of a specific threat from the enemy). ACEVAL trials were flown in the air combat maneuvering instrumentation range, which was contained in a circle of 34 nautical miles northwest of Nellis Air Force Base, Nevada, and was designed to track, monitor, and record up to eight high-performance aircraft simultaneously.

For the primary measures of effectiveness, or the dependent variables, the test design used "loss rate," or the ratio of one side "killed" in a test trial to the total number on that side at the beginning of the trial, and "exchange ratio," or the ratio of enemy forces killed to friendly forces killed. The fundamental measure of encounter outcome was limited to the number of aircraft losses on each side. Secondary measures of effectiveness were included for events during air combat, such as detecting and identifying the opposing force and firing weapons.

There were two separate experiments, each consisting of 360 valid trials of mock combat between a Blue or friendly force of F-14A or F-15A aircraft and a Red or enemy force of F-5E aircraft simulating the Soviet MIG-21J. The test design called for varying numbers of trials for each combination of encounter size, force ratio, and GCI condition. Aircraft differing from the test aircraft--the Navy's TA-4 and A-7 and the Air Force's A-7, F-4, and F-100--were used in 44 "intruder" trials to simulate a neutral third force and enforce the test requirement that the enemy be identified visually before weapons could be fired.

Friendly aircraft from the Air Force were flown by eight F-15 pilots; the Navy's friendly force consisted of six F-14 pilots and six F-14 naval flight officers. The enemy force was made up of nine Air Force pilots, six Navy pilots (one of whom resigned early in ACEVAL, leaving five), and one Marine pilot. Six GCI controllers were assigned to the friendly forces and six were assigned to the enemy. All the test participants represented the most highly skilled in the armed services.

In addition to the primary control variables, test control operating instructions and rules of engagement were developed, in an attempt to insure a consistent and reliable data base. These procedures governed, among other conditions, knowledge about the opposition's flight composition and tactics, trial start and separation distances, visual identification before firing, and natural trial conclusions (when every aircraft in one force had either been killed or disengaged by departing from the test area into a designated "safe" area).

Tactics were not controlled other than through the requirement that each force comply with its own tactical doctrine, amended for the threat of "all-aspect" weapons. (An all-aspect weapon can be fired at a target that is within the shooter's range, regardless of what part of the target it is facing. It is unlike infrared missiles that require the shooter to maneuver to the rear of a target before firing.) The forces were also obliged to observe safety constraints, the test-control operating instructions, and the rules of engagement. However, the tactics that were used were documented for later analysis. Aircrew assignments were made by squadron leaders, but in flight the aircrews had complete freedom over when and where to use their weapons.

The scenario for all ACEVAL trials was a fighter sweep mission. One force was to clear a given area of the opposition, which in turn was to intercept the aircraft that had penetrated its area. The friendly aircraft had aids--the television sight unit on the F-14 and the Eagle Eye II on the F-15 for visually identifying targets at greater distances than would otherwise be possible. The enemy aircraft had no visual aids but had a radar homing and warning system that was simulated by verbal calls. The following weapons were simulated:

	F-14A	F-15A	F-5E
Missiles			
Radar (AIM-7F)	4	4	0
Infrared (AIM-9L)	4	4	4
	Offboresight	Offboresight	Boresight
Ammunition rounds			
M-61 Gatling gun (20 mm)	682	940	0
Cannon (23 mm)	0	0	200

Aircrews were required to identify opponents visually before firing. However, the rules of engagement required only that the first missile firing on each side be at a visually identified target; all firings thereafter could be without visual identification.

ACEVAL did not follow the normal progression of events in most test programs. Its planning, initial instrumentation validation, and participant training were preceded by AIMVAL, which gave ACEVAL's managers and participants considerable experience

in a test similar to ACEVAL. AIMVAL provided them extended "training" and "pretesting." Moreover, special training missions for ACEVAL were flown during the last portion of AIMVAL.

F-14 ACEVAL test trials began on June 2 and ended on October 19, 1977; F-15 trials began on June 24 and were finished on November 10. The normal daily sequence of missions was to fly the friendly against the enemy aircraft in combinations of 4v4, 2v4, and 1v2, with the takeoffs staggered to provide variations. The last two missions, starting late in August, were combined into one 4v4 mission.

Overall, 70.9 percent of all the scheduled missions were flown, for a total of 1,119 trials. Weather and range instrumentation problems were the reasons for most of the cancellations. Of all the attempted trials, 398, or 35.6 percent, were judged invalid, mostly because of data omissions, absence of aircrew tallies, violations of safety, and the weather.

The JTF placed considerable emphasis on the accuracy, continuity, and consistency of the data. The main tasks of quality control were making checks on the manual and automated data collection systems; formally and systematically reviewing each trial for validity as it occurred, including in the data base only valid trials, and looking for biases from the exclusion of invalid trials. The JTF also controlled the distribution of data in order to provide a common data base for analyses. Trial validations were given to the Navy, the Air Force, and the Institute for Defense Analyses; information on invalid trials was provided when it was available.

The JTF's reports on ACEVAL present the analysis that had been completed by 60 days after the test flying was finished. The analysis benefitted from the continuing presence of the aircrews, who helped interpret the quantitative results; the analysts, who observed the test trials, were able to report their awareness of idiosyncrasies in the data. The JTF analyzed several measures of effectiveness representing process and outcomes, used several analytical tools (including histograms, frequency counts, analysis of variance, contingency tables, and multiple linear regression) to determine trends and relationships in the data, used operational (rather than observed) loss rates in regression analysis to compensate for the effects of the random number of generator that was used to determine kills during test trials, normalized the data for an equal number of trials in each GCI condition, tested various relationships, and tested for significance in order to determine the strength of the statistical relationships that had been determined.

One of the original purposes of ACEVAL had been to obtain empirical data for one-on-one air combat that could be used to help predict the outcome of larger air battles. The availability of the air combat maneuvering instrumentation range in 1976 had made it

possible to run operational tests and evaluations of aircraft encounters up to 4v4. Accordingly, in its ACEVAL reports, the JTF reported that it had addressed this issue by determining whether 1v1 data could be used to predict 4v4 outcomes, examining the applicability of models based on the Lanchester theory of combat (attrition models that attempt to describe the effects of the concentration of fire power by means of differential equations), and attempting to develop different models using multivariate analysis.

The JTF's conclusion was that 1v1 exchange ratios are largely irrelevant to the exchange ratios of large force ratios--that is, knowing how an aircraft performs in one-on-one combat does not provide a measure of its performance when there are more combatants or when it is outnumbered. Further, attempting to predict larger air combat outcomes from smaller ones by means of the attrition models was inappropriate, regardless of the data, because the ACEVAL results were from small, discrete engagements whereas the models are meant to predict large, continuous engagements. Finally, since the ACEVAL data were found to fit several different models that yielded different results, trying to address this issue for ACEVAL would have been inappropriate, given that there was no agreed-upon theoretical framework for choosing the best analytical model.

Other difficulties about addressing this issue of predicting or "extrapolating" from smaller engagements to larger ones were pointed out by the aircrews. One of the Air Force pilots, for example, observed that a 2v1 fight differs fundamentally from a 20v10 fight because the greater number of aircraft leads to more errors in perception. It might be possible for pilots who are current in air combat and tactical intercept training to maintain total awareness of a but not a larger fight of Also, ACEVAL was not designed for comparing the outcomes ot combat between aircraft of equal capability or for determining the structure of fighter forces. Further, the friendly and enemy forces were not configured so as to represent either current or future combat between U.S. and Soviet forces. The threat force had advanced weapon capabilities projected for 1985 but lacked current radar missile and radar warning and homing capabilities; the friendly forces had weapons currently available for close air combat but were aided by advanced radar and visual devices. Thus, ACEVAL's loss rates and other outcomes have little applicability beyond the specific test aircraft.

Finally, the aircrews had knowledge that pilots do not have in combat, where they cannot be sure that they have a complete tactical picture and must assume that they are outnumbered. Aircrews fight differently when they do not know the number of aircraft or there are too many to keep track of, but in ACEVAL the number was small and they knew what it was. As the Navy On-Site Analysis Team pointed out, "you can't extrapolate from small

numbers of fighters operating in isolation to larger numbers of aircraft (fighters and others) attempting to achieve or prevent something in a complex environment" (II.B.20, p. 1).

ACEVAL was undertaken not only to provide empirical data about air combat but also to establish a test methodology that might be useful beyond this initial test in an operational air combat test program. The JTF assessed the test's constraints and procedures to determine, if possible, the effect of the test on the data and to make recommendations for conducting future tests. About the operational constraints, the JTF concluded that they did not bias the results but that only the data trends would be useful in trying to project the results to other conflicts. This is partly because the constraints of environment, instrumentation, safety, and the like, although discussed in the test plan and controlled for by various operating instructions and rules of engagement, were not varied or measured for their actual effect. The observations that were made about their probable effect were subjective and open to dispute by others with different experiences and opinions. Even the usefulness of the data trends is uncertain, since these too might have been different if the constraints had been modified.

As for the effectiveness of the test procedures, the JTF supported its judgments about the problems that occurred during planning, implementation, analysis, and reporting with exclusively subjective evidence, or what it called "a free and open expression of the success and difficulties experienced by the Joint Test Force" (II.B.17, p. iii). No criteria for assessing how "effectiveness" would be demonstrated were proposed in the test feasibility study, design, or plan. Nevertheless, at the conclusion of the test, the JTF made a systematic review of how ACEVAL had been conducted and put considerable effort into documenting the "testing" lessons that had been learned, offering possible alternatives for future tests.

ITEM 3: Air combat maneuvering instrumentation

The air combat maneuvering instrumentation system at Nellis Air Force Base was designed to track, monitor, and record up to eight high-performance aircraft simultaneously. It used simultaneous measurements from several ground stations for "real time" computation that determined the position of each aircraft with respect to the ground references. Aircraft attitudes (their orientation in relation to their direction of motion) were determined from data communicated from the aircraft through an integral data link to a situation display at the control center. These data, including particulars of the engagement, were recorded on magnetic tape and used for later briefings.

The system at Nellis had been modified for ACEVAL so that it could detect and track eight simulated missile or gun firings simultaneously and determine their targets. After a process of weapon-simulation validation, the JTF ascertained that it was

possible to get a 90 percent agreement between what the system said about success in intercepting targets and what the missile and gun simulations showed. However, instrumentation problems during AIMVAL led to a remechanization of the system to improve its infrared missile target-to-background discrimination. The remechanized system examined a missile's self-tracking seeker for 1.2 seconds (0.7 seconds longer than in AIMVAL). At the end of that time, a failure would be scored and attributed to the remechanization if the seeker's "look" angles had diverged from the target detected by the system by more than 2.0° (compared with 6.9° in AIMVAL). Tone falling below the missile's tracing threshold of 36-38 decibels (greater than AIMVAL's 20) was also failure.

All this favored the friendly forces in three ways. First, the enemy F-5E carried only one AIM-9 captive test unit while the friendly F-14 and F-15 carried two. This meant that the friendly aircrews could switch to the second unit for a second shot of the missile after a delay of only 1.3 seconds, whereas the enemy aircrews had a 3- to 4-second delay before the system could recover for their second shot. Second, the way in which the seek angles were measured for enemy aircraft made them more susceptible to "jumps" than the F-14 or F-15 were, so that the enemy force had a higher percentage of failures attributable to remechanization. The result was a tendency to invalidate trials with large numbers of enemy firings, which in turn may mean that the data are either disproportionately high or disproportionately low for AIM-9 activity. Third, the rise in the minimum tone for infrared tracking corrected for invalid lock-ons but prevented what would otherwise have been valid shots at 20 to 30 decibels. The rise in tone threshold gave the friendly forces more time to employ the AIM-7.

ITEM 4: Trial validation

The JTF placed considerable emphasis on data accuracy, continuity, and consistency. It attempted to control the quality of the data related to critical trial events, the status of aircraft and instrument operations, meteorological conditions, and the qualifications, qualitative assessments, and operational conditions of the aircrews. Extensive and thorough quality assurance checks on automated and manual data collection were incorporated into the data acquisition process in an attempt to minimize original collection, transcription, and computation errors. From information on the quality of the data on the test trial events, each trial was validated to make sure that essential instruments and weapons had been functional, that the test procedures had been complied with from start to finish, and that the data were sufficient and accurate.

This process of validation was accomplished by a committee of three, who represented JTF operations, JTF data management, and the Institute for Defense Analyses. Advisory members represented JTF engineering and test control and the Air Force and Navy analysis teams. Additionally, the aircrews were allowed to provide written rebuttals to the committee's statements, which were

evaluations of the quality and continuity of the data rather than judgments about the operational realism of the test trials.

The committee reviewed each day's trials the next day. Decisions deferred for lack of information were made as soon as possible after it was available. Each "fire event" had to be documented with a known outcome, which was evaluated for the accuracy of the reasons for calling it a "kill" or a "miss." To salvage trials with incorrect or undetermined outcomes but otherwise sufficient data, a post-trial simulation was conducted to determine a possible outcome. When more than one such simulation was made for a trial, the results of only the first were accepted, in order to preclude data manipulation. The committee attempted to insure that these post-trial results were consistent with other events. The validation procedures and criteria that the committee used in ACEVAL had emerged throughout AIMVAL. They were defined and itemized in an effort to make judgments consistent. The committee accepted all trials as valid until proven otherwise. A trial was invalid if critical data were inaccurate or missing or if the test procedures had not been followed. It was called "no trial" if the range equipment or software had been inoperable.

Since validation was essentially a screening process, and the JTF was concerned about bias in favor of the trials that had had less activity, the JTF examined the committee's validations for validation rates for the various encounter sizes, the post-trial simulation outcomes, and the omission of "fire events" and "kills." Regarding the overall validation rates, the JTF found that they did not change with encounter size except for the 4v4 trials. Approximately 44 percent of the 4v4 trials were declared invalid; for the other encounter sizes, approximately 30 percent were invalid. As for the aircraft, 36 percent of the F-15 trials were invalid, with 29 percent for the F-14. Many trials were unavoidably invalid because of gross malfunctions in the system or the absence of visual detection on both sides. Of the remaining invalid trials for which data were collected and usable, the exchange ratio was better for the enemy forces. Thus, it appears that if the validation process introduced any overall bias, it favored the friendly forces. Looking specifically at the 4v4 trials, for example, reveals that friendly loss rates were higher and exchange ratios were lower for both the F-14 and F-15.

The JTF determined that the post-trial simulation of outcomes did not significantly affect the test results because the outcomes differed greatly from "real time" outcomes for only 11 (or 1.5 percent) of the 720 trials that were compared. Outcomes for 93 trials (48 F-14 and 45 F-15) were determined by the post-trial simulations, of which 77 were validated; in the correct missile result was called a miss, with no effect on the trial or the data base. For the that were kills (one was not accounted for in the JTF analysis), were earlier kills in real time and thus had no effect on the results. The remaining 11 trials may be biased but were scattered among 42 test trial bins, none containing more than 2 post-trial simulations.

Since only one visual detection was required for trial validity, the ACEVAL data base included trials with no fire events (21 F-14 and 28 F-15); trials with fire events but no kills, one or both forces successfully disengaging and leaving the area (61 F-14 and 75 F-15); and trials in which some or all participants on either side were killed. Overall, 103, or 28.6 percent, of the F-15 valid trials had no kills; there were 82, or 22.8 percent, valid F-14 trials with no kills. The absence of fire and kills prevailed in the lower encounter sizes. The lack of engagements is a possible, natural outcome of air combat encounters but may be exaggerated in the ACEVAL data because of test design constraints. The ACEVAL data reflect no engagements for almost one fourth of the encounter outcomes.

ITEM 5: Overall loss rates and exchange ratios

ITEM 6: The effect of disengagement on kill and loss rates

	Disengaged			
	F-14		F-15	
	Yes	No	Yes	No
Friendly				
Kill rate				
Loss rate				
Enemy				
Kill rate				
Loss rate				

ITEM 7: Modeling problems

The simulation models that were used in ACEVAL for determining the probability of kill did not take into account the vulnerability of the F-14 and F-15 as targets. Tables that were set up for each combination of weapon and target contained data for the characteristics of the weapons (missile fusing, warhead, and the like) but the data on target vulnerability were taken from earlier testing, except that the vulnerability of the enemy F-5E was based on information about the MIG-21. Since vulnerability models for the F-14 and F-15 were not available in 1976, models of a substitute, the F-4, were used for both aircraft. The JTF did not analyze or discuss critical differences between the characteristics and performance of the F-4 and the F-14 and F-15, and the results that were reported do not reveal how the trial outcomes were affected by them. The JTF speculated that

The simulation models that were used in ACEVAL for the AIM-9 missile included the throttle setting for "military power" but were incapable of determining whether an aircraft was working at any lower setting. During the test, however, aircraft often used "idle power" against head-on adversaries, in order to reduce the infrared signature and, hence, the range at which infrared weapons could be locked on and launched. The use of idle power had not been anticipated, and it was not monitored by the test instrumentation or included in the simulation models. Nevertheless, the JTF indicated, without citing evidence, that the modeling problem affected friendly and enemy forces equally.

A radar clutter model was added, possibly incorrectly, to the basic real-time AIM-7F simulation to make it more representative for firings at lower altitudes. The AIM-7 simulation did not properly account for long-range "lookdown" clutter effects that lowered the probability of a missile hit, nor did it account for clutter constraints such as beam crossover for missile altitudes higher than 15,000 feet, so that the results may be optimistic. Successful crossover at lower than 15,000 feet was allowed within the last 0.1 seconds of flight, but this may have yielded pessimistic results. The JTF indicated, without evidence, that these problems favored the friendly forces.

ITEM 8: Missile activity against targets

	AIM-7 trials		AIM-9 trials		AIM-9 trials	
	F-14	F-15	F-14	F-5E	F-15	F-5E
No. of attempts to fire						
No. of target intercepts						
Interceptions divided by attempts						
No. of target kills						
Kills divided by attempts						

ITEM 9: F-15 learning as expressed in exchange ratios

F-15 trials	Exchange ratio	
	First half	Second half
All		
All, with no new pilots		
All, with at least one new pilot		

ITEM 10: Reported effects of weather on F-14 kill ratios by force ratio

	1v1	1v2	4v4
Trials without good weather			
Kill ratio			
Without good weather			
With good weather			
Overall			
Relation to overcast			

ITEM 11: Other uses of ACEVAL's results

The JTF expressed cautions about how the results should be used, given ACEVAL's technical inadequacies, but this has not precluded their use in a number of unintended ways.

ACEVAL's results have been quoted often, along with AIMVAL's, in the public debate on whether to build up U.S. military defenses with quantity or quality. Interpreting aspects of the ACEVAL data selectively to support one position or the other is inappropriate, because the test was not designed to address this issue.

The Rand Corporation analyzed the implications of ACEVAL's results for future air-to-air combat and for acquisition policy for fighter aircraft for the Air Force. Rand made no attempt to extrapolate to scenarios or conditions not examined in ACEVAL but did use the data on the differences in the performance of the F-14 and F-15 during long-range combat to recommend procuring a "high-

low mix" of aircraft: a "high" component of aircraft equipped to detect, sort, identify, and attack enemy aircraft at long range and a "low" component of smaller and less costly aircraft equipped for close encounters. Rand's inferences about performance are reasonable but the conclusion that there is a need for this combination of aircraft cannot be derived from the test results. Long-range and short-range air combat occurred during ACEVAL trials primarily because of an imbalance in friendly and enemy equipment and instrumentation.

OSD organizations gave funds to Stanford Research Institute International and the Institute for Defense Analyses to develop generic models with which to extrapolate ACEVAL and AIMVAL data. MACEVAL is an algorithmic set of one-dimensional, scalar rules for modeling individual air combat engagements and accumulating the statistics in ACEVAL-like aggregations of exchange ratios and loss rates as a function of engagement size. The model was used to support the view that ACEVAL's results reflect general principles beyond the specifics of the test. MACE is a mathematical attrition model for predicting the outcome of close air combat between groups of aircraft within visual identification range. The model was said to be able to reproduce the outcome of ACEVAL's close encounters up to 4v4 by using only the parameters of the 1v1 encounters. Such generalizations should be treated with caution, since the ACEVAL data reflect only its design and its implementation.

The BDM Corporation, under contract with the Air Force, examined the utility of adding a laser weapon system to the F-15 in the ACEVAL/AIMVAL environment. After constructing its own model, BDM concluded that an automated antimissile laser device would significantly reduce the F-15's losses. The Air Force Tactical Fighter Weapons Center contracted with VEDA, Inc., to analyze the ACEVAL data on the F-15 to determine how the results differ when AMRAAM parameters are substitued for AIM-7 parameters. (AMRAAM is a medium-range air-to-air missile made by Hughes Aircraft.) The conclusion was that all the measures of performance showed that AMRAAM would give the F-15 greater engagement opportunity than the AIM-7. The two efforts at modeling are similarly misleading because of the specificity of the tactics, countertactics, use of weapons, and rules of engagement to the ACEVAL test as it was flown, any one of which might have been different had any aspect of the test been different. ACEVAL's aircrews were very competitive and took advantage of hardware, software, psychology, and everything else at their disposal to "win" that test, with the result that much of the data are unique to the test rather than generalizable to combat.

In another study of ACEVAL's tactics and training, BDM demonstrated that both influenced the test's results. Attempting to develop insight for the use of "high" aircraft performance and its effects on aircraft survivability and lethality, the Institute for Defense Analyses found that ACEVAL's aircrews tended to use all the thrust their aircraft had, used speed brakes extensively, and

seemed to be more successful as aircraft use rates rose. Such process data are important, especially for developing tactics and training, but caveats similar to those above on the possible distortion of the data must not be overlooked. Both the Navy and the Air Force indicated that their practical experiences in ACEVAL had been useful for their air combat operations while, at the same time, they were very outspoken about the test's limitations.

Finally, the Air Force Tactical Weapons Center used the JTF's analysis of ACEVAL's implementation to design the simulation tests for the evaluation of AMRAAM's operational utility, overcoming several of the inadequacies in ACEVAL's design in doing so. The AMRAAM simulation was less costly than ACEVAL, unconstrained by safety factors, included a variety of mission scenarios, and controlled for poor weather, communications jamming, and other parameters not tested in ACEVAL. Nevertheless, ACEVAL results (for example, the distribution of ranges for first radar detections) were used in the simulation program. Since these data are highly dependent on the particularities of ACEVAL's implementation, the simulation's results are probably overly optimistic.

(973542)

END

FILMED

5-84

DTIC